Genus-level studies of gene dynamics for the Aspergillus genus

Theobald, Sebastian

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Genus-level studies of gene dynamics for the *Aspergillus* genus

PhD thesis by
Sebastian Theobald

Technical University of Denmark

January 2018

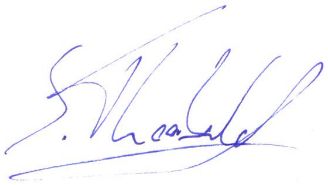
Contents

Contents

Preface

This thesis serves as a partial fulfillment of the requirements to obtain a PhD degree of the Technical University of Denmark (DTU). The work was carried out at the Department of Biotechnology and Biomedicine under the supervision of Prof. Mikael R. Andersen, Assistant Prof. Tammi C. Vesth, Prof. Thomas O. Larsen and Anders G. Pedersen. The project has been funded by the Villum foundation (grant VKR023437).

Kongens Lyngby, January 2018



Sebastian Theobald

Acknowledgements

This project would have not been possible without the many wonderful people involved who supported me professionally and personally. First, I would like to thank my supervisors Mikael Rørdam Andersen and Tammi Camilla Vesth for giving me the opportunity to pursue this project and mentoring me during the three years. You were putting me on the right path when I was getting lost in the details of the project — which in my case happens quite frequently. I learned so much during these three years. Thank you for making this possible and also teaching me that there's much more to science than just the project. Furthermore, I would like to thank the whole Network Engineering of Eukaryotic Cell Factories group. Especially Jane Nybo, the best office mate you can wish for, with constant positive and energetic attitude, Inge Kjærbølling for also being full of positive energy and curiosity, and Julian Brandl for his humor and supporting me with python. It was a great experience to work and travel together with you guys to so many different places.

This mostly bioinformatic project tried to bridge the gap between comparative genomics and molecular biology, which would have not been possible without the help of my colleagues. I would like to thank Jakob B. Hoof and Jakob K. Rendsvig for a great collaboration to elucidate the malformin gene cluster and Thomas Ostenfeld Larsen, who joined as co-supervisor later in the process, for his advice and mentoring regarding the chemical aspects of this thesis. Also I would like to thank Matthias Brock for the opportunity to visit his laboratory at University of Nottingham and Elena Geib for collaboration. Here also a big thanks to Justyna Kuska! It was great to have a small masters reunion in the UK!

DTU Bioengineering has been a great place to work and I want to thank everyone in building 223 and 221 for creating such a good atmosphere. I'm very grateful that I could work together with so many inspiring people during my PhD — I definitely learned a lot from everyone.

Also, I would like to thank my family and friends for their constant support they gave me during these past three years. Finally, a very special thanks goes to Ângela de Carvalho for the love, the endless support, and for just always being there.

Abstract

The fungal genus *Aspergillus* has a great impact on society. It includes species with industrial value *A. niger* and *A. oryzae*, medically relevant species *Aspergillus fumigatus* and *A. terreus*, and the model organisms *A. nidulans*.

A. nidulans, chosen as model organism due to its sexual cycle and possibility to study genetics using spore color, provided insights into key regulators of cell cycle control, development and secondary metabolism. Results obtained from studies on *A. nidulans* were found to be applicable on other *Aspergilli* as well.

Genome sequencing of *A. nidulans*, *A. fumigatus* and *A. oryzae* further revealed a large phylogenetic distance between species of the genus. Hence, conclusions derived from studies on model organisms might not apply to the genus as a whole but rather on smaller taxonomical groups. Taxonomy of Aspergilli is constantly under evaluation and the information added by new methods and technologies is aiding the definition of new taxonomic groups (Chen et al., 2016; Hubka et al., 2016).

In the 300 *Aspergillus* genome project, we are extending the set of sequenced species to further investigate genome and gene diversity across the entire genus. This specific project deals with the diversity of secondary metabolite (SM) genes and conservation of regulators. To achieve a comparison of several species at once, we created methods which classify genes into families. This approach highlighted that regulators like *mcrA*, a master regulator of secondary metabolism, is conserved throughout Aspergilli and that *galR*, a regulator which was thought to be unique for *A. nidulans* is present in many other species. SM genes needed a slightly different type of clustering since they show high similarity between each other — making them difficult to separate. The full pathway for a secondary metabolite is encoded on a spatial collective of genes, a secondary metabolite gene cluster (SMGC). Thus, to find families of SMGC for production of similar SMs we compared whole gene clusters against each other. Our analysis highlights differences of SMGC families shared on several taxonomic levels.

Characterizing regulators and SMGC over several species as families has the advantage that we are able to follow their distribution throughout the genus. This has several benefits. Examination of conserved genes can give insights into the adaptations of a section of Aspergilli. Examining SMGC families can give hints to uninvestigated SMGC producing promising drug leads. Overall, analyzing *Aspergillus* species with this comparative approach will reveal their gene dynamics and diversity and give insights into adaptations of sections throughout the genus.

Dansk resume

Den filamentøse svampeslægt *Aspergillus* har stor indflydelse påsamfundet. Slægten indeholder arter med stor industriel værdi (f.eks. *A. niger* og *A. oryzae*), medicinsk indflydelse (f.eks. *A. fumigatus* og *A. terreus*), og biologisk model organismer (f.eks. *A. niger* og *A. nidulans*). *A. nidulans* blev en klassisk model organisme grundet sin seksuelle cyklus og muligheden for at studere genetik ved hjælp af sporefarve. Dette har givet indsigt i essentielle regulatorer af cellecykluskontrol, udvikling og sekundær metabolisme. Studier foretaget i *A. nidulans* har vist sig at være anvendelige i flere af arterne fra *Aspergillus* slægten. I 2005 blev *A. nidulans*, *A. fumigatus* og *A. oryzae* genom sekventeret og dette afslørede en langt større fylogenetisk afstand mellem arterne end før antaget. Derfor er det vigtigt ikke at antage at konklusioner fundet i studier af model organismer kan overføres direkte påslægten som helhed, men snarere påmindre taksonomiske grupper. Taxonomien for *Aspergillus* slægten er konstant under evaluering, hvor information opnået via nye metoder og teknologier understøtter definitionen af nye taxonomiske grupper (Chen et al., 2016; Hubka et al., 2016). Vi vil, igennem *Aspergillus* hel-slægts sekventeringsprojektet, udvide sættet af sekventeret arter for yderligere at undersøge genom og gendiversiten over hele slægten. Denne afhandling omhandler diversiteten af sekundære metabolit (SM) gener og konserveringen af regulatorer. For at sammenligne flere arter påé n gang, har vi skabt metoder der grupperer gener i familier af samme funktion eller produkt. Disse metoder har fremhævet at regulatorer som *mcrA*, en masterregulator af sekundær metabolisme, er bevaret i hele *Aspergillus* slægten og at *galR*, en regulator der blev anset for at være unik i *A. nidulans* er til stede i langt flere arter. SM gener bruger en speciel type gruppering, da de viser høj genetisk similaritet mellem hinanden — hvilket gør dem vanskelige at adskille. Den fulde pathway af en sekundær metabolit er kodet i et spatial kollektiv af gener, et sekundært metabolit genkluster (SMGC). For at finde familier af SMGC til produktion af tæt beslægtede SM'er har vi sammenlignet hele genkluster mod hinanden. Vores analyser viser forskelle og ligheder mellem SMGC-familier, der deles påflere taksonomiske niveauer. Karakterisering og gruppering af regulatorer og SMGC over flere arter har det fortrin, at vi er i stand til at identificerer deres tilstedeværelse og følge deres fordeling over hele slægten. Dette har flere fordele. Undersøgelser af konserveret gener kan give indsigt i evolution og tilpasning af sektioner i *Aspergillus* slægten. Undersøgelser af SMGC-familier kan lede til nye SMGC der potentielt producerer lovende lægemidler. Samlet set vil analyser af *Aspergillus* arter med denne sammenlignings metode afslører arternes gendynamik og diversitet, samt give indsigt i tilpasninger af sektioner i hele

slægten.

Publications

Book chapter included in this thesis

- Approaches for Comparative Genomics in *Aspergillus* and *Penicillium* Nybo, J.L.*, Theobald, S.*, Brandl, J., Vesth, T.C. Andersen, M.R., Aspergillus and Penicillium in the Post-genomic Era. de Vries, R.P., Benoit Gelber, I. & Andersen, M.R. eds., June 2016, Caister Academic Press, Ch. 4.

Submitted manuscripts

- The genomes of *Aspergillus* section *Nigri* reveal drivers in fungal speciation
  Vesth, T.C., Nybo, J.L., Theobald S., Frisvad, J.C., Larsen, T.O., Nielsen, K.F., Hoof, J.B.,Brandl, J., Salamov, A., Riley, R., Nielsen, M.T., Lyhne, E. K., Kogle, M. E., Strasser, K.,McDonnell E.,Barry, K., Clum, A.,Chen C., Nolan, M., Sandor, L., Kuo, A., Lipzen, A., Hainaut, M., Drula, E., Tsang, A., Henrissat B., Wiebenga, A., Mäkelä, M.R., de Vries R.P., Grigoriev, I.V., Mortensen, U.H., Baker, S.E.*, Andersen, M.R.*

- Uncovering bioactive compounds in *Aspergillus* section *Nigri* by genetic dereplication using secondary metabolite gene cluster networks
  Theobald, S., Vesth, T., Rendsvig, J.K., Nielsen, K.F.,Riley, R., de Abreu, L.M., Asaf Salamov, Jens Christian Frisvad, Thomas Ostenfeld Larsen, Mikael Rørdam Andersen*, Jakob Blæsbjerg Hoof*

Manuscripts in preparation

- Genus level analysis of PKS-NRPS and NRPS-PKS hybrids reveals their origin in Aspergilli
  Theobald, S., Vesth, T.C., Andersen, M.R.

- Comparative genomics of *Aspergillus nidulans* and section *Nidulantes*
  Theobald, S., Vesth, T.C., Nybo, J.L., Kjærbølling, I.,Riley, R., Salamov, A., Lyhne, E. K., Kogle, M.E., Frisvad, J.C., Hoof, J.B., Mortensen U.H.,Dyer, P., Momany, M., Larsen, T., O., Baker, S., Andersen, M.R.

Manuscripts in preparation and not included in this thesis

- Comparative genomics and investigation of melanin biosynthesis pathways in *Aspergillus* section *Terrei*

Theobald, S., Vesth, T.C., Nybo, J.L., Geib E., Kjærbølling, I.,Riley, R., Salamov, A., Lyhne, E. K., Kogle, M.E., Frisvad, J.C., Hoof, J.B., Brock, M., Mortensen U.H., Larsen, T. O., , Baker, S., Andersen, M.R.

- Comparative genomics in *Aspergillus* sections *Usti* and *Cavernicolus* Nybo, J.L., Vesth T.C, Theobald, S., Frisvad J.C, Larsen T.O., Kjærbølling I., Lyhne, E.K., Kogle M.E., Phippen C., Barry K., Clum A., Chen C., Nolan M., Sandor L., Lipzen, A., Kuo A., Salamov A.A., Drula E., Riley R., Henrissat B., Hainaut M., Wiebenga A., Mäkelä M.R., de Vries R.P., Grigoriev I.V., Baker S.E.*, Andersen M.R.*

Thesis aim and structure

The aim of this thesis is to characterize gene dynamics on a genus level. Furthermore, a new method to characterize secondary metabolite genes had to be established. This thesis describes a new method to characterize spatial collectives of genes for production of secondary metabolites (SMs): secondary metabolite gene clusters (SMGCs), over several genomes. *Aspergillus* section *Nigri* and *Nidulantes* are used as case studies to:

1. Classify SMGC into SMGC families producing similar SMs
2. Characterize the SMGC diversity on different taxonomic levels.
3. Provide gene leads for elucidation by combining presence of SMGC families with metabolite production data.
4. Characterize SMGC families using a guilt by association approach.
5. Show the phylogenetic history of one SM class
6. Identify distributions of regulators throughout the genus

First, the introduction provides an overview of the importance of fungal biology, secondary metabolites, *Aspergillus* genome research and approaches for comparative genomics in *Aspergillus* and *Penicillium*. Chapter 2 includes the developed method for SMGC family creation and addresses points 1,2 and 3. Chapter 3 builds up on the preceeding chapter, extends the analysis performed under point 2 and adds point 4. Chapter 4 extends the dataset by more species outside section *Nigri* and shows the evolution of a hybrid SM class. Chapter 5 illustrates the general applicability of the method by using section *Nidulantes*. Points 1-4 and 6 are addressed here.

Chapter 1

Introduction

1.1 Fungi and their impact on society

The genus *Aspergillus*, a taxon of filamentous fungi within the *Ascomycetes*, is of vast importance in different areas of society. In nature, they act mostly as plant degraders and to cope with the complex plant cell wall components, they secrete a diverse repertoire of hemicellulases, pectinases and cellulases. These enzymes are useful in industrial applications and many of these are produced by the important cell factory *Aspergillus niger* CBS 513.88 (de Vries et al., 2001). In food-related applications, e.g. juice production, the enzymes degrade cell wall components that would usually make the food processing inefficient. Fungal enzymes are used to ease the process and maximize yield at low cost. Furthermore, fungal xylanases and endoglucanases are used to separate gluten and starch from flour (Moore et al., 2011), thus creating the desired product. Another isolate of *A. niger*, ATCC 1015, is the parent strain of the strains used in industry in the production of e.g. citric acid (E330), a food preservative and flavor enhancer, and gluconic acid, a food additive and acidity regulator. Another fungus, *A. oryzae*, is heavily used in the food industry for the production of e.g. soy sauce and miso.

Apart from their industrially useful enzymes, Aspergilli also produce secondary metabolites (SMs) which can be both problematic and useful compounds. *A. flavus* contaminates stored food stuffs with strong carcinogen aflatoxin (Klich, 2007). Aflatoxin is then ingested by consumers where it causes liver cancer and suppresses the immune system. It can also be transferred from mothers to newborns, creating an even greater health threat(Rafiei et al., 2014). *Aspergillus* species of section *Nigri* affect food stocks like e.g. cashew nuts where contaminating *Nigri* species produce fumonisins, ochratoxin A, and secalonic acid (Lamboni et al., 2016). These

contaminations can affect the economy. For instance, cashew nuts are an important export product of west African countries, which in the case of Benin made up 8% of export revenues in 2011 (Lamboni et al., 2016).

SMs are also found throughout other *Ascomycete* species than Aspergilli. *Cochliobolus heterostrophus* (or: *Bipolaris maydis*) is another *Ascomycete* from the class of *Pleosporaceae* and causes corn leaf blight in maize. Yang (1996) identified a protein responsible for the necrotrophic effector T-toxin, a SM, which affects T-cytoplasm corn — a male sterile variant corn. *Magnaporthe grisea*, a *Sordariamycete* produces a SM, through the protein Ace1, that is an avirulence factor triggering resistance in rice (*Oryza sativa*) plants which carry the resistance gene Pi33 (Böhnert et al., 2004). Ace1 is only expressed during plant infection; any disturbance of the process abolishes Ace1 expression (Böhnert et al., 2004).

Despite the toxic and virulence promoting SMs, many can be used as pharmaceuticals, e.g. to treat bacterial and fungal infections. *Penicillium chrysogenum* (now identified as *P. rubens*) for example, has been long used for the production of the antibiotic penicillin — which has also been discovered in *A. nidulans* (van Liempt et al., 1989). Another SM isolated from fungi, lovastatin (Rubinstein et al., 1991), has been widely used as a cholesterol-lowering drug and might have the same impact on society as penicillins. Other compounds are investigated for their pharmaceutical properties like e.g. cytochalasins (Petersen et al., 2014b) and malformins (Hagimori et al., 2007b). Hence, it is of interest to investigate the enzymatic and genetic basis for SMs.

SMs are synthesized by different classes of enzymes. In fungi, these are polyketide synthases (PKS), non-ribosomal peptide synthetases (NRPS), enzymes only consisting of a smaller subset of modules (PKS-Likes, NRPS-Likes), fusions of PKS and NRPS (PKS-NRPS/ NRPS-PKS hybrids) terpene cyclases (TC), and dimethylallyl transferases (DMATS). These enzymes produce a molecule that is further modified by tailoring enzymes which e.g. add additional groups or cyclize the initial molecule. The enzymes necessary for production of a SM are encoded by a collective of genes — a secondary metabolite gene cluster (SMGC). This thesis concentrates mostly on PKS, NRPS and derivates of these SM genes. Sequencing of fungal genomes facilitated to reveal the genetic basis for many SMGC. With the first *Aspergillus* species sequenced in 2005 (Galagan et al., 2005), some SMGC could be elucidated. The genetic basis for the majority of SMs, however, still remained hidden since genome mining for SMGCs is laborious. Additionally, the evolution of SMGC, leading to analogous compounds, was impossible to uncover because only few Aspergilli were sequenced.

1.2 Non ribosomal peptide synthetases (NRPS)

Non-ribosomal peptides (NRPs) constitute a major group of secondary metabolites, like e.g. the anticancer compound nidulanin (Klitgaard et al., 2015), anticancer drug enhancing malformins (Hagimori et al., 2007a), the iron-transporting siderophores (Eisendle et al., 2003), antifungal echinocandins (Bills et al., 2016), and antimicrobial penicillins (Wright, 1999). The latter was first discovered in *P. chrysogenum* (Diez et al., 1990), and shortly after, under certain fermentation conditions, in *A. nidulans* (Holt and MacDonald, 1968; MacCabe et al., 1991).

This wide array of bioactive NRPs is synthesized by non-ribosomal peptide synthetases NRPSs. These enzymes can — as opposed to ribosomal peptide synthesis — utilize L-amino acids, as well as non-proteogenic amino acids as their substrate (Finking and Marahiel, 2004), hence creating compound diversity. They consist of one or more modules, which typically contain three domains (Weissman, 2015; Marahiel, 2016); an adenylation (A) domain for loading of amino acids, a peptidyl carrier protein (PCP) domain, also called thiolation (T) domain for peptide chain transfer, and a condensation (C) domain for peptide bond formation. Other domains like the epimerisation (E) domain change the chirality of their proximate amino acid. A-domains choose the amino acid according to the Stachelhaus-code — a sequence inside the A-domain in their active site (Stachelhaus et al., 1999). First, amino acids are recognized and activated by the A-domain, and react with ATP to aminoacyl-adenosinmonophasphate (aminoacyl-AMP) which can then form a aminoacyl thioester with the panthetheine group of the T-domain (a process analogous to activation of amino acid for ribosomal peptide synthesis) (Evans, 2016). Condensation domains show different subtypes that are dependent on the stereoisomer of the preceding amino acid: starter C domain as initial C domain, C^{LL} domain for joining two L-amino acid, C^{DL} domains for joining of D and L configured amino acids — selecting the correct enantiomer (Clugston et al., 2003)— and heterocyclizing C domains (Finking and Marahiel, 2004). Epimerising C domains (E) are a special subtype. These domains switch the stereochemistry of an amino acid from L to D. Dual epimerisation-condensation domains show both functions. Surprisingly, examples of NRPS following different synthesis schemes are increasing. Enniatin, enterobactin, bacillibactin and gramicidin S are synthesized by iterative NRPS, meaning that their modules are reused (Mootz et al., 2002).

Penicillin is produced from joining L-aminoadipic acid — a product of L-lysine modification — with L-cysteine and L-valine (Fig. 1.1). Each amino acid is recruited by the adenylation domain and loaded onto the T domain. L-aminoadipic acid and cystein are condensed and subsequently, L-valine is

isomerized by the E-domain to its D form and the C-domain condenses D-valine with the dipeptide. The thioesterase domain releases the tripeptide which is further processed by tailoring enzymes to form penicillin. In general, the release of peptides is either by hydrolisis, leading to a linear peptide, or lactamization, leading to a cyclic peptide (e.g. nidulanin (Klitgaard et al., 2015)).

Malformin, previously considered as antibiotic against bacteria (Suda and Curtis, 1966), has been shown to be a cytotoxin which arrests cells in G2 and M phase and subsequently induces cell death (Wang et al., 2015a). Cancer cells only repair DNA damage in G2 stage (Hartwell and Kastan, 1994), hence this is a potential anticancer drug enhancer. Malformin is a pentapeptide of Val, D-Leu, Ile and two D-Cys which contains a disuflide bridge between the two cysteins. Its NRPS has not been elucidated until now (see chapter 3). Nidulanin A, another NRP, is a cyclic tetrapeptide consisting of L-phenylalanine (Phe), L-valine (Val) and D-Val, one L-kynurenine residue (a metabolite of L-Trp) and isoprene (Klitgaard et al., 2015). These compounds emphasize that NRPS in fungi show a wide variety of structures, incorporating various monomers and epimerizing amino acids. The applicability of non ribosomal peptides is shown by the function of penicillin and malformin. The investigation of secondary metabolism in *Aspergillus* can benefit society with candidates for pharmaceutical compounds.

Bioinformatic work has greatly assisted in the investigation of NRPS domains and diversity. Known examples of NRPS have been used to create prediction tools for adenylation domain specificity, like e.g. NRPSpredictor (Rausch et al., 2005). Due to their conservation, these domains have also been the primary subject for phylogenetic and phylogenomic studies. Phylogenomic studies identified nine NRPS subfamilies throughout 38 fungal genomes (Bushley and Turgeon, 2010). Several subgroups are suspected to be of bacterial origin, emphasizing that horizontal gene transfer (HGT) from bacteria to fungi is common (Bushley and Turgeon, 2010). Another key finding was that a group of multimodular NRPS is restricted to fungi and extensive gain and loss of domains drives NRPS evolution. Additionally, NRPSs drive chemical innovation through point mutations, rearrangements, duplication and deletion of domains/modules, creating an even higher chemical diversity through incorporation of new substrates (Fischbach et al., 2008). Other domains which have been investigated are the C-domains. Phylogenetic analysis of C-domains revealed a common origin of several classes (Rausch et al., 2007).

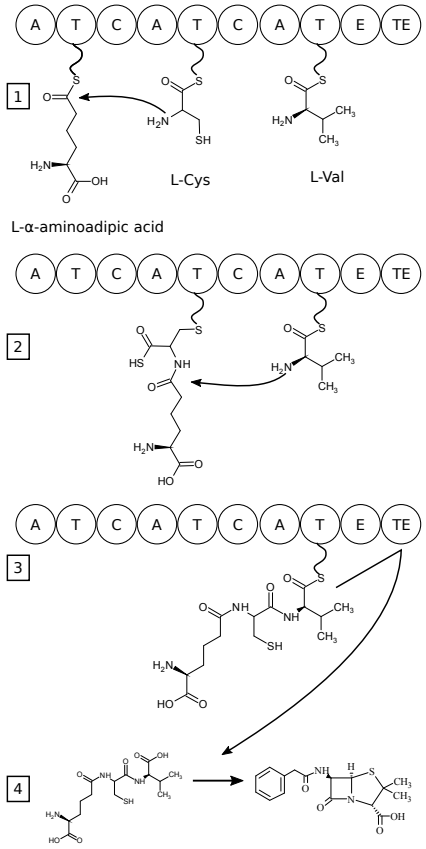


Figure 1.1: Synthesis of δ-L-α-aminoadipyl-L-cysteinyl-D-valine, a penicillin precursor, by an NRPS synthetase. **1**: L-cys-thioester condenses with L-α-aminoadipyl-thioester to form a dipeptide. **2** L-Val is epimerized to D-Val and condenses with the dipeptide. **3** δ-L-α-aminoadipyl-L-cysteinyl-D-valine (ACV) is released. **4** tailoring enzymes process ACV to penicillin G

1.3 Polyketide synthases (PKS)

Polyketides, the most abundant secondary metabolites in *Aspergillus* species, are a class of diverse compounds with a wide range of bioactivities. These compounds include mycotoxins such as sterigmatocystin (Purchase and Van der Watt, 1973) and the highly carcinogenic aflatoxin (Coulombe and Sharma, 1985), while others are insecticides as in the case of dehydroaustinal (Valiante et al., 2017). They can also have medical properties like the cholesterol-lowering drug lovastatin (Rubinstein et al., 1991).

Polylketide synthases consist of several domains to synthesize these compounds. In the first step of polyketide synthesis, the acyl transferase (AT) domain loads acetyl-coenzyme A (acetyl-CoA) on the pantetheine arm of

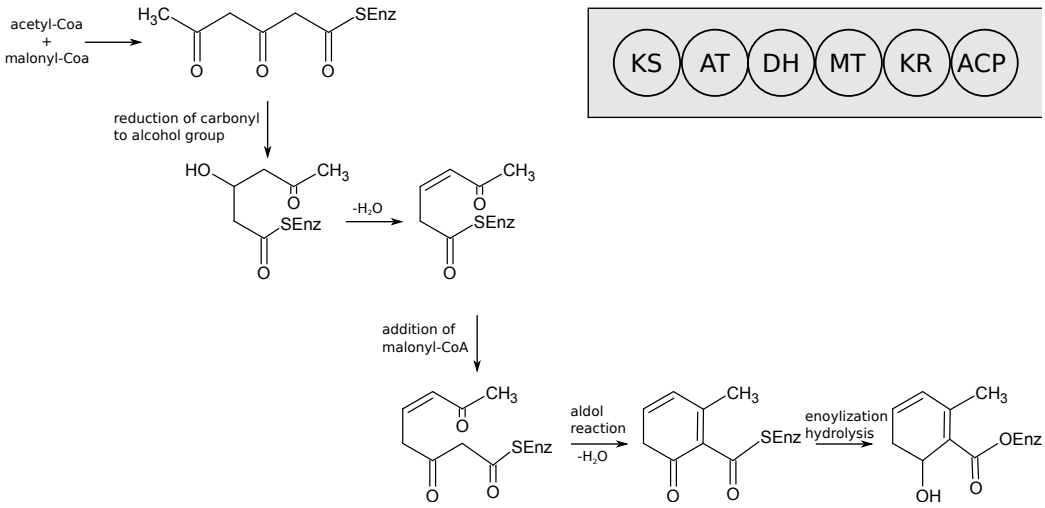


Figure 1.2: Synthesis of 6-methylsalicilic acid and structure of PKS. The 6-msa synthetase is a good example of a minimal fungal iterative type I PKS and illustrates the functions of reducing domains. Adapted from (Dewick, 2009)

the ACP domain which then transfers acetyl-CoA to the β-ketoacylsynthase (KS) domain. The extender unit malonyl-coenzyme A (malonyl-CoA) is loaded onto the ACP domain and the KS domain condenses acetyl-CoA and malonyl-CoA to yield the growing polyketide chain. Subsequently, ketoreductase (KR), dehydratase (DH), enoylreductase (ER) and methyltransferase (MT) domains can further reduce the polyketide. Depending on their β-keto processing activity the final polyketide chain can be highly reduced (HR), partially reduced (PR) resulting in macrolides, and non-reduced (NR), which will results in aromatics — classifying the PKS. The KR domain reduces the β-ketoester to a hydroxyester, DH domain dehydrates to an conjugated ester while the enoyl reductase will further reduce it to a reduced ester. The extending and processing steps are repeated until the polyketide has reached its final length, is loaded on the thioesterase (TE) domain, and released.

The 6-methylsalicyllic acid (6-MSA) synthase is a good example of a minimal fungal iterative type I PKS (Fig. 1.2). Acetyl-CoA and two malonyl-CoA are condensed to a poly-β ketoester, which is subsequently reduced and dehydrated to form a double bond. After condensation of another malonyl-CoA an aldol and enolization reaction further modify the ketoester until it is hydrolized and released from the enzyme as 6-MSA. The 6-MSA synthetase is a proposed progenitor of other SMs, like e.g. patulin and yanuthone D, especially in the genus *Aspergillus* (Frisvad and Larsen, 2015).

Phylogenomic analysis of KS domains indicated eight groups of fungal PKSs (Kroken et al., 2003). One ancestral PKS is orsellinic acid synthase (orsA), which has undergone a vast duplication event in *Pezizomycotina* (Koczyk et al., 2015). This event is responsible for the amount of non-reducing PKSs we can identify in fungi. From the orsillinic acid synthase, meroterpenoid synthases, naphthopyrone synthases and aflatoxin synthases (among others) developed, shaping the chemical diversity we see in fungi.

1.4 Polyketide - non ribosomal peptide synthetase hybrids

The enzymes classes producing polyketide and non-ribosomal peptides — PKSs and NRPSs — can also be fused into a PKS-NRPS or NRPS-PKS hybrid enzymes, creating a chimeric compound. Most identified fungal hybrids have a PKS-NRPS orientation, the PKS part being highly reducing (Boettger and Hertweck, 2013). There are however, a few fungal NRPS-PKS, e.g. tenuazonic acid (Yun et al., 2015). The products of hybrids are e.g. the mycotoxin cyclopiazonic acid (Sorenson et al., 1984), the antioxidant pyranonigrin (Miyake et al., 2007; Awakawa et al., 2013) and the potential anti-cancer compound cytochalasin (Petersen et al., 2014a).

For cytochalasin biosynthesis in *Aspergillus clavatus*, an octaketide is formed by condensation of acetyl-CoA and 8 malonyl-CoA units and, after three methylations, processed by Diels-Alder cyclization (Boettger and Hertweck, 2013; Qiao et al., 2011). Subsequently, the octaketide is joined with phenylalanine to create an amide. After release, it is further modified by tailoring enzymes.

The hybrid gene, *ccsA*, for cytochalasin biosynthesis has an interesting evolutionary history. In a study by Khaldi and Wolfe (2011), they showed evidence that *ccsA* is actually a *syn2* homolog which was horizontally transferred from *Magnaporthe grisea* to *A. clavatus*. *syn2* is a duplication of *ace1* we covered in section 1.1. In general, this hybrid seemed to undergo massive duplications but also deletions, since it was, at the time, not found in other *Aspergillus* species (Khaldi and Wolfe, 2011). Other studies suggest an interkingdom transfer of a NRPS-PKS hybrid (Lawrence et al., 2011). We further investigate the role of the hybrid SM class, also regarding the *ccsA* hybrid, in chapter 3.

1.5 Genome research in Aspergilli

The first genome sequences of *A. nidulans* (Galagan et al., 2005), *A. fumigatus* (Nierman et al., 2005), and *A. oryzae* (Machida et al., 2005) provided insights into the phylogeny, evolution and diversity of these fungi. Predicted orthologs between *A. nidulans* and *A. fumigatus* shared 66% amino acid identity (Galagan et al., 2005) — which is comparable to the protein identity shared between mammals and fish (Dujon et al., 2004). More *Aspergillus* species were sequenced later; among them *A. niger* (Pel et al., 2007) which provided insights into primary metabolism as well as secondary metabolism genes. The sequencing and comparison of *A. niger* ATCC 1015 to the sequenced CBS 513.88 was the first study to explore differences and comparative genomics between two closely related species of section *Nigri* (Andersen et al., 2011). The study highlighted the great differences between isolates with a chromosomal arm inversion, high mutation rates, and differences in SMGC content identified in *A. niger* ATCC 1015.

The availability of *Aspergillus* genome sequences also increased the amount of SMs that could be linked to their SMGCs. Since analogous SMs, compounds of slightly different structure, are detected throughout the genus (Frisvad and Larsen, 2015), studies were searching for similar SMGCs. The basic local alignment search tool (BLAST) (Camacho et al., 2009) can be used individually on certain species to find candidate clusters. Maximum likelihood phylogenies are an alternative to relate secondary metabolite genes and explain their phylogeny. They are limited however to conserved domains rather than the whole cluster to estimate similarity between SM enzymes. Multiple domains per enzyme make it difficult to handle in data post processing and data trimming as well as the used algorithm affect the outcome of phylogenetic analysis (Zhou et al., 2017). Additionally, different types of SMs have to be treated with individual analyses and with increasing amount of sequences getting reliable branches through bootstrapping can be a problem. Projects like the 1000 fungal genomes project, the *Fusarium* sequencing projects, or our *Aspergillus* sequencing project (extending to over 300 species) require a genome mining method that highlights SMGC dynamics in several species at once.

Instead of creating individual comparisons of SMGCs on species to species basis or phylogenetic trees covering each class of SM gene, we created a tool for categorization of SMGCs into SMGC families — a group of SMGCs producing similar compounds. To achieve this, we created all against all protein BLAST which extracts bidirectional best hits related to SMGC proteins. The scores of BLAST hits are then aggregated to create a SMGC to SMGC similarity network. Random walks (Pons and Latapy, 2005) are used on the net-

work to yield families of SMGCs. We show how SMGC families can be used to aid in the characterization of SMGCs — replacing laborious blast comparisons and synteny analyses and providing strong leads for linking genes to metabolites. As a case study, we chose the SM rich genus *Aspergillus* to show the applicability of our method (see chapter 2,3, and 5). Furthermore, we show how our method can highlight SMGC diversity on different taxonomic levels.

1.6 Approaches for Comparative Genomics in *Aspergillus* and *Penicillium*

Approaches for Comparative Genomics in *Aspergillus* and *Penicillium*

4

Jane L. Nybo,* Sebastian Theobald,* Julian Brandl, Tammi C. Vesth and Mikael R. Andersen

Abstract

The number of available genomes in the closely related fungal genera *Aspergillus* and *Penicillium* is rapidly increasing. At the time of writing, the genomes of 62 species are available, and an even higher number is being prepared. Fungal comparative genomics is thus becoming steadily more powerful and applicable for many types of studies.

In this chapter, we provide an overview of the state-of-the-art of comparative genomics in these fungi, along with recommended methods. The chapter describes databases for fungal comparative genomics. Based on experience, we suggest strategies for multiple types of comparative genomics, ranging from analysis of single genes, over gene clusters and CaZymes to genome-scale comparative genomics. Furthermore, we have examined published comparative genomics papers to summarize the preferred bioinformatic methods and parameters for a given type of analysis, highly useful for new fungal geneticists. Moreover, the chapter contains a detailed overview of comparative genomics studies of key fungal traits such as primary metabolism, secondary metabolism, and secretome analysis.

Finally, we gaze into a possible future of the field by comparing the current state of fungal comparative genomics to the development in bacterial genomics, where the comparison of hundreds of genomes has been performed for a while.

Introduction

The advent of the genomic era for the *Aspergillus/Penicillium* group started in 2005 with a set of three publications in Nature covering the most studied organisms in the genus, *Aspergillus nidulans*, *A. oryzae* and *A. fumigatus* (Galagan *et al.*, 2005; Machida *et al.*, 2005; Nierman *et al.*, 2005). These analyses gave us the first genome-scale overview of the three individual species, but notably, a large part of all three manuscripts was dedicated to the comparison of the genomes to each other or other fungal genomes (e.g. *Saccharomyces cerevisiae*). It was already clear from these first studies that a large part of the understanding of the individual species and genetic features comes from *comparative genomics* in addition to the initial genome analysis. The comparative analysis gave new information about a variety of features such as mating factor variations; primary and secondary metabolism; adaptation to different environments; genome dynamics; conserved non-coding sequence; and pathogenicity. Similarly, the later publication of the genome sequence for *Penicillium chrysogenum* (van den

*These authors contributed equally.

Berg *et al.*, 2008) based the analyses of phylogeny and chromosome dynamics on comparisons with the previously published aspergilli.

Since then, the interest in comparative genomics has only increased. A notable application is in the strength of comparative genomics to infer annotations from well-studied model organisms, such as the ones mentioned above, to less studied species, which are being sequenced as part of current large genome sequencing efforts in both Penicillia and Aspergilli (see 'Current status of genomics', below). Furthermore, genome mining of fungi using tailored algorithms for comparative genomics has long been interesting due to the wealth of natural products found in these species (Bok *et al.*, 2006).

In this chapter, we define *comparative genomics* as *the comparison of genetic features for large parts or whole genomes* and will focus on applications and studies of this type. The comparison of single genes across organisms is clearly a highly valuable effect of the availability of genome sequences, but will not be addressed here.

It is clear from the examples presented above, and what will be shown in the following sections, that there is a high analytical power in comparing features across genomes. In this context, we wish to provide overviews of the following:

- the current status of fungal genomics in *Aspergillus* and *Penicillium*;
- the available tools and databases for fungal comparative genomics, both general and specialized applications;
- results and potentials of comparative genomics studies within the field.

Furthermore, we will present the main application areas of comparative genomics, as well as the current best practice for comparative genomics within the individual research areas. Our goal is to enable the reader to identify the current best practices for fungal comparative genomics based on an overview of algorithms and preferred parameters in the field.

Current status of genomics

The foundation of comparative genomics is the availability of relevant genome sequences. Currently (May 2015), 62 genome sequences from the *Aspergillus* (32 species) and *Penicillium* (30 species) genera are available from public repositories (Table 4.1). While the majority of the sequences are from individual species, an increasing amount of sequences are addressing multiple strains of the same species. In those cases, the primary application of the genome has been comparative genomics with the objective to identify strain-specific traits. Examples include high pencillin production in *P. chrysogenum* NCPC10086 (Wang *et al.*, 2014), studies of pathogenicity and secondary metabolism in *A. fumigatus* Af293 versus A1163 (Fedorova *et al.*, 2008; Sanchez *et al.*, 2012), and enzyme-producing phenotypes in *A. niger* CBS 513.88 (Andersen *et al.*, 2011). Details regarding these cases, as well as cross-species comparative genomics will be discussed in the following sections.

It is quite clear given the current number of multiple independent genome sequencing efforts, as well as the JGI Community Sequencing Programs, that the number of available *Aspergillus* and *Penicillium* genome sequences will increase rapidly in the near future. Current ongoing efforts include both re-sequencing of multiple isolates of the same species (for instance clinical isolates, basic research as for the *S. cerevisiae* 100 genomes project (Strope *et al.*, 2015), or results of industrial strain improvement programs) as well as a large number

Table 4.1 List of *Aspergillus* and *Penicillium* species and strains with publicly available genome sequences

Species	Strain	Reference
A. acidus	CBS 106.47	[1]
A. aculeatus	ATCC16872	
A. brasiliensis	CBS 101740	[1]
A. campestris	CBS 348.81/IBT28561	[3]
A. carbonarius	ITEM 5010	
A. clavatus	NRRL1	Fedorova *et al*., 2008
A. flavus	NRRL3357	Nierman *et al*., 2015a
A. fischerianus/N. fischeri		Fedorova *et al*., 2008
A. fumigatus	A1163	Fedorova *et al*., 2008
A. fumigatus	Af293	Nierman *et al*., 2005
A. glaucus	CBS 516.65	[1]
A. kawachii	IFO 4308	Futagami *et al*., 2011
A. nidulans	FGSC A4	Galagan *et al*., 2005
A. niger	ATCC 1015	Andersen *et al*., 2011
A. niger	CBS 513.88	Pel *et al*., 2007
A. niger	NRRL3	
A. niger van Tieghem	ATCC 13496	
A. novofumigatus	CBS117520/IBT16806	[3]
A. ochraceoroseus	CBS 550.77/IBT245754	[3]
A. oryzae	RIB40	Machida *et al*., 2005
A. parasiticus	SU-1	Linz *et al*., 2014
A. phoenicis	ATCC 13157	
A. rambelli	SRRC1468	
A. ruber/Eurotium rubrum	CBS 135680	Kis-Papo *et al*., 2014
A. steynii	CBS112812/IBT 23906	[3]
A. sydowii	CBS 593.65	[1]
A. terreus	NIH 2624	
A. tubingensis	CBS 134.48	[1]
A. ustus	3.3904	Pi *et al*., 2015
A. versicolor	CBS 583.65	[1]
A. wentii	DTO 134E9	[1]
A. zonatus		[1]
P. aethiopicum	IBT 5753	
P. aurantiogriseum	NRRL 62431	Yang *et al*., 2014
P. bilaiae	ATCC 10455	[2]
P. brevicompactum	AgRF18	[2]
P. camemberti	FM 013	Cheeseman *et al*., 2014
P. canescens	ATCC 10419	[2]
P. capsulatum	ATCC 48735	

Table 4.1 Continued

Species	Strain	Reference
P. chrysogenum	NCPC10086	Wang *et al*., 2014
P. digitatum	Pd1	Marcet-Houben *et al*., 2012
P. digitatum	PHI26	Marcet-Houben *et al*., 2012
P. expansum	ATCC 24692	[2]
P. expansum	R19	Yu *et al*., 2014
P. expansum		Li *et al*., 2015
P. expansum	PEXP	Ballester *et al*., 2015
P. expansum	PEX1	Ballester *et al*., 2015
P. expansum	PEX2	Ballester *et al*., 2015
P. fellatanum	ATCC 48694	[2]
P. glabrum	DAOM 239074	[2]
P. italicum		Li *et al*., 2015
P. italicum	PITC	Ballester *et al*., 2015
P. janthinellum	ATCC 10455	[2]
P. lanosocoeruleum	ATCC 48919	[2]
P. marneffei	ATCC18224	Nierman *et al*., 2015b
P. nordicum	BFE487	
P. oxalicium (*decumbens*)	114-2/CGMCC 5302	Liu *et al*., 2013a
P. paxilli	ATCC 26601	Berry *et al*., 2015
P. raistrickii	ATCC 10490	[2]
P. roqueforti	FM164	Cheeseman *et al*., 2014
P. rubens[4]	Wisconsin 54-1255	van den Berg *et al*., 2008
P. stipitatum/Talaromyces stipitatus	ATCC 10500	Nierman *et al*., 2015b

[1] Genomes sequenced as a part of a JGI community sequencing proposal (CSP) led by Ronald P. de Vries (CBS-KNAW, NL). Publication pending.
[2]Genomes sequenced as a part of a JGI CSP led by Dave Greenshields (Novozymes).
[3]Genomes sequenced by JGI as a part of a JBEI/DTU sequencing proposal.
[4]Species previously thought to be *P. chrysogenum.*

of *de novo* sequencing projects. Currently, it seems likely that the number of *Penicillium* and *Aspergillus* genome sequences combined will exceed 100 in 2016. Such a development renders traditional storage and analysis methods inefficient and increases the need for databases, which can make these sequences available to the research community, as well as provide tools for comparing the wealth of genomes.

Available genome databases with comparative genomics capabilities

The fungal research community has been fortunate in the availability of several data repositories dedicated to analysis of fungal genomes or specific features thereof. The original source of inspiration for many of these efforts has been the immensely successful Saccharomyces

Genome Database (SGD) (Cherry *et al.*, 2012), which formed the basis for the framework behind the Aspergillus Genome Database (AspGD) (Arnaud *et al.*, 2010; Cerqueira *et al.*, 2014). Since the launch of AspGD, several new databases and repositories have been made available, each adding specific analysis features and/or additional genomes relevant for comparisons. Table 4.2 gives an overview of fungal and general databases currently available which give access to fungal genomes and offer various types of comparative genomics tools.

Table 4.2 denotes which databases offer which types of services, but of course all tools are not identical even with similar scopes, and have different strengths. All databases furthermore have individual design philosophies and are tailored towards different communities. For the following types of comparative genomics studies based on information available in databases, we recommend the following approaches:

- *Investigation of orthologues of single genes.* An efficient approach would be to look up the gene in FungiDB, AspGD or the JGI webpages, if the gene/species of interest is present in any of these, and from there navigate to the gene orthology feature of each webpage. Alternatively, one can employ the BLAST interface present at many of the sites (Camacho *et al.*, 2009). To get a comprehensive overview including as many organisms as possible, use both the JGI portal or NCBI GenBank as these contain genomes not found elsewhere.
- *Analysis of gene clusters across organisms.* Currently, the simplest approach would be to navigate to the JGI secondary metabolism feature, identify the gene cluster of interest, and from there navigate to the JGI genome browser. Here, one can examine gene conservation in a few related species. For a more sophisticated analysis, where the gene cluster is known, one can use FungiDB or AspGD. Both databases offer detailed synteny mapping based on the SYBIL software (Crabtree *et al.*, 2007) to a larger number of organisms. This allows the user to examine whether gene order has been rearranged.
- *Analysis of CAZymes.* For both *Aspergillus* and *Penicillium* species, this is a highly interesting analysis of potential biomass-degrading enzymes. Such data are currently best examined by downloading it from the CAZy-database (Cantarel *et al.*, 2009), if it is available for the organisms of choice, and performing an analysis offline using custom comparative tools. Alternatively, the staff behind CAZy are involved in many community efforts of this type.
- *Untargeted analysis.* In some cases, it is of interest not to look at defined/known genes, but instead compare genes of a specific type, structure or function across organisms. For this type of analysis, FungiDB/EupathDB have highly sophisticated gene search methods, based on DNA and protein motifs, domains, signal peptides, association with specific metabolites, etc.
- *Genome-scale comparative genomics.* At the moment, this type of analysis is not available through any of the platforms. This is currently best undertaken by downloading genome data of the organisms of choice and performing these computationally heavy tasks on dedicated systems.
- *Download of genomics data sets.* Depending on the genomes of interest, NCBI Genomes or JGI are the best sources of data. Some genomes are only found in one database but not the other, so for a comprehensive analysis, both databases must be queried. In our hands, we find that for most of our applications, the data format downloadable from JGI is the most flexible, as the sequence data is preformatted in several different ways. Furthermore,

Table 4.2 Overview of often-used resources for fungal comparative genomics

		Comparative genomics tools											
Database	Reference(s)	BLAST	Synteny analysis	Custom analysis flows	Orthologue identification	Secondary Metabolite Cluster comparison	Visual comparison of genomes	Comp. Genome Browser	Comparison to other omics	Comp. Gene Locus information	Cross-genome searches for annotation	Metabolic Pathway Analysis	Bulk Data Download
AspGD	Arnaud *et al.*, 2010; Cerqueira *et al.*, 2014	•	•		•			•	•	•	•		•
CADRE	Gilsenan *et al.*, 2012												•
CAZy	Cantarel *et al.*, 2009										•		•
EuPathDB	Aurrecoechea *et al.*, 2013	•	•	•	•			•	•	•	•		•
FungiDB	Stajich *et al.*, 2012	•	•	•	•			•	•	•	•	•	•
JGI genome portal	Grigoriev *et al.*, 2012	•			•	•		•	•[1]	•			•
MycoCosm	Grigoriev *et al.*, 2014	•			•	•		•	•[1]	•			•
NCBI	Wheeler *et al.*, 2013	•											•

[1]EST/SRN-sequencing data are integrated as a genome browser track on many genomes.

the JGI has a specialized interface (GLOBUS) for downloading large datasets, which is very useful for studies involving the comparison of more than a few species.

For larger or more advanced, or specialized studies, it is typically advantageous to use some of the many specialized algorithms and methods developed. The next section gives an overview of these.

General methods for comparative genomics

Methods in comparative genomics are in general based on tools used for single genome analysis, and in many cases need to be tailored accordingly to be able to manage cross-species differences. Comparative genomics has been used to shed light on everything from identifying species-specific genes, transcription factor (TF) binding sites, repeat elements, and single nucleotide polymorphisms (SNPs) to predicting phylogenetic trees, genome scale models, secretomes, and horizontal gene transfers. These approaches can generally be sorted into two broad categories: Studies identifying orthologues/paralogues (see 'Identification of orthologues across species', below) and studies identifying synteny across genomes (see 'Whole genome comparisons and synteny analysis', below).

Selecting the right method for a specific analysis depends on multiple factors, the most important ones being the research topic or question and the type of data involved. The data type and quality can have a large impact on the strength of the analysis and a variety of methods have been developed to deal with items such as variations in annotation and sequencing quality. Because of these large variations and biological questions, the right tools and methods should be used in the right context. The following section describes different cases and the methods used, and presents an overview table of studies and types of analysis (Table 4.3).

Identification of orthologues across species

As is evident from Table 4.3, there is a wide range of applications, methods, and parameters for identifying orthologues. A key component of all these methods is to compare either DNA or protein sequences. Sequence alignment is a method to compare the DNA, RNA or proteins to identify similar regions. There are a large number of alignment algorithms addressing scalability, speed, and parameter optimization, where some are mentioned in Table 4.3.

The predominant alignment algorithm for comparative genomics is the Basic Local Alignment Search Tool (BLAST), which combines database searches with a fast heuristic Smith–Waterman-based local alignment approach (Altschul *et al.*, 1990). It is perhaps the most widely used tool, but does come with some pitfalls, which warrants a detailed discussion. A common orthologue selection approach is to use bidirectional (reciprocal) best hits between genomes. It is a very rigorous approach that almost certainly will find true orthologues (Wolf and Koonin, 2012), but will miss as much as 60% of potential orthologues (Dalquen and Dessimoz, 2013), in that only one hit is identified for each gene or protein. Another, less stringent approach is to use reciprocal hits in combination with additional selection criteria, among other *expectancy value* (*e-value*), *alignment identity* (%) and *alignment coverage* (%), which increases the likelihood of identifying orthologous relations and allows the identification of multiple orthologues for a given gene or protein. In general, one

Table 4.3 Overview of methods and parameters applied in comparative genomics studies of penicillia and aspergilli

	Agren *et al*., 2013	Andersen *et al*., 2008	Andersen *et al*., 2011	Arnaud *et al*., 2010	Ballester *et al*., 2015	Braaksma *et al*., 2010	Coutinho *et al*., 2009	David *et al*., 2008	Fedorova *et al*., 2008	Flipphi *et al*., 2009	Galagan *et al*. 2005	Gibbons and Rokas, 2009	Inglis *et al*., 2013	Kis-Papo *et al*., 2014	G. Liu *et al*., 2013	J. Liu *et al*., 2013	Machida *et al*., 2005	Marcet-Houben *et al*., 2012	Pel *et al*., 2007	Pi *et al*., 2015	Rokas *et al*., 2007	van den Berg *et al*., 2008	Vongsangnak *et al*., 2008	Wang *et al*., 2014	Toolbox	Visualization	Online	Local installation	Database
Genus																												
Penicillium	×				×									×	×			×				×		×				
Aspergillus		×	×	×		×	×	×	×	×	×	×	×			×	×		×	×	×		×					
Type of study																												
Genome-scale metabolic model	×	×	×					×								×			×									
Database				×																								
Genome sequenced	×		×		×		×		×		×			×	×		×	×	×			×		×				
Specialized comp. genomics										×		×																
Focus																												
Strain	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×				
Clade/group										×	×									×								
Core					×				×	×	×			×	×			×	×	×	×		×	×				
Homology/orthology																												
BLAST alignment																												
Bidirectional[6] best hits	×	×				×			×	×				×	×	×	×	×	×	×	×		×					
Modified mutual best hit				×																								
BLASTp	×	×		×	×	×	×	×	×	×	×			×	×	×	×	×	×	×	×		×					
E-value	$>1e^{-30}$				$>1e^{-05}$		$>1e^{-30}$	$>1e^{-50}$	$>1e^{-05}$	$>1e^{-30}$					$>1e^{-10}$	$>1e^{-30}$		$>1e^{-05}$	$>1e^{-10}$	$>1e^{-10}$	$>1e^{-05}$							
Bit score																	>100											
Alignment identity	>40%					>40%					>80%				>40%				>50%				>40%					

Alignment coverage	>50%			>50%									>60%			>50%		>50%													
Alignment coverage of shortest hit					>70%																										
Alignment length	>200aa																>100aa					>200aa									
Sequence length similarity					>80%																										
tBLAST *Bit score*														>100													x	x	x		
BLASTx *AL*[II] ≥ *25%* *AI*[III] ≥ *50 aa*																						x					x	x	x		
BLASTn *E–value*		$1e^{-75}$	x																								x	x	x		
Alignment-based clustering																															
InParanoid			x									x															x	x		x	
TRIBE-MCL													x																x		
OrthoMCL														x													x	x	x	x	
E-value									$>1e^{-10}$				$>1e^{-05}$																		
Match cutoff									40%				>50%																		
CAFE[IV]													x																x		
Strain/clade/core-specific feaure identification																															
Synteny																															
Sybil			x					x		x									x					x	x	x				
BLAT																							x			x	x				
MUMmer								x										x	x	x					x		x				
PASA pipeline								x											x						x		x				
Single nucleotide polymorphism																															
SAMtools[V]															x									x		x	x			
BWA[V, VI]				x																							x				
MUSCLE[V, VII]															x										x	x				
GATK[VIII]				x																				x			x				
SOAP[IX]																							x	x			x				

Table 4.3 Continued

	Non-coding or repeats	Repeat-Masker	RepeatScout	Tandem Repeats Finder	Transposon-PSI	tBLASTn	Emboss etandem[X]	Multi-lagan	RSAT[XI]	Cosmo[XII]	TRANSFAC[XIII]	Secondary metabolism	SMURF[XIV]	BLASTp	*E-value*	*Alignment identity*	*Alignment coverage*	*Alignment coverage (query)*
Database			×	×							×							
Local installation		×	×	×	×	×		×	×	×	×		×	×				
Online		×		×		×	×	×	×		×		×	×				
Visualization							×		×		×			×				
Toolbox		×						×	×	×	×							
Wang *et al*., 2014																		
Vongsangnak *et al*., 2008																		
van den Berg *et al*., 2008		×	×	×	×								×					
Rokas *et al*., 2007																		
Pi *et al*., 2015													×	×	$>1e^{-10}$	>50%		>50%
Pel *et al*., 2007																		
Marcet-Houben *et al*., 2012																		
Machida *et al*., 2005																		
J. Liu *et al*., 2013																		
G. Liu *et al*., 2013													×					
Kis-Papo *et al*., 2014																		
Inglis *et al*., 2013													×					
Gibbons and Rokas, 2009							×											
Galagan *et al*. 2005								×										
Flipphi *et al*., 2009									×	×	×							
Fedorova *et al*., 2008		×	×	×	×	×							×					
David *et al*., 2008																		
Coutinho *et al*., 2009																		
Braaksma *et al*., 2010																		
Ballester *et al*., 2015													×	×	$>1e^{-05}$		>50%	
Arnaud *et al*., 2010																		
Andersen *et al*., 2011																		
Andersen *et al*., 2008										×								
Agren *et al*., 2013																		

Present when % genes in cluster was found			>50%																												
Secretome																															
SignalIP[XV]	-			>0.43						x		x					x	x										x	x	
PSORT[XVI]	x													x			x	x										x	x	
TMHMM[XVII]										x		x					x											x	x	
Phylogeny																															
Idenification of conserved genes																															
BLASTp			x			x					x			x			x		x								x	x	x	
E-value			>1e^{-05}			>1e^{-04}					>1e^{-10}			>1e^{-05}			>1e^{-10}													
Alignment identity																	>80%													
Alignment coverage			>50%								>60%			>50%			>80%													
Identical number of introns						x													x											
Similar length						x													>95%											
Alignment																															
GeneDoc							x											x									x		x	
ClustalW aa		x				x					x				x	x											x	x	x	
ClustalX aa					x						x		x		x												x	x	x	
MAFFT			x							x				x													x	x	x	
MUSCLE [DNA/aa]			x			x								x					x								x	x	x	
Kalign			x																								x	x	x	
DIALIGN-TX														x														x	x	
M-Coffee[XVIII]			x											x													x	x	x	
Evolutionary clustering																															
Gblocks										x	x																x	x	x	
RaxML			x							x				x														x	x	
PhyML/BioNJ[XIX]			x											x												x		x	x	

Table 4.3 Continued

	MEGA[XIX]	Phylip[XX]	TreeCon[XIX]	Bootstrap ML[XII]	BLASTp for HGT	*Cluster absent in closely related species*	*Cluster found in distant related species*	*Cluster sequence identity*	Visualization	ETE toolkit[XXII]	FigTree[XXIII]
Database											
Local installation	×	×	×			×	×	×			×
Online						×	×	×		×	
Visualization	×	×	×			×	×	×		×	×
Toolbox	×	×		-						×	
Wang *et al*., 2014											
Vongsangnak *et al*., 2008											
van den Berg *et al*., 2008		×									
Rokas *et al*., 2007											
Pi *et al*., 2015	×			1000		×	×	>50%			
Pel *et al*., 2007	×			1000							
Marcet-Houben *et al*., 2012				100		×	×			×	
Machida *et al*., 2005											
J. Liu *et al*., 2013											
G. Liu *et al*., 2013	×			1000							×
Kis-Papo *et al*., 2014											
Inglis *et al*., 2013											
Gibbons and Rokas, 2009				×							
Galagan *et al*. 2005		×		1000							
Flipphi *et al*., 2009	×			1000							
Fedorova *et al*., 2008		×									
David *et al*., 2008											
Coutinho *et al*., 2009				1000							
Braaksma *et al*., 2010											
Ballester *et al*., 2015				100						×	
Arnaud *et al*., 2010											
Andersen *et al*., 2011			×	1000							
Andersen *et al*., 2008											
Agren *et al*., 2013	-										

Functional annotation

GO			x	x	x				x	x			x	x	x				x	x				x	x	x	x	x	x	x	x	x	x	x
Pfam	x							x	x		x	x		x	x			x	x				x				x	x	x	x	x	x	x	x
AntiSMASH													x							x							x	x	x	x	x	x	x	x
InterPro	x		x		x								x	x	x			x	x					x			x	x	x	x	x	x	x	x
KEGG	x	x	x		x					x				x	x	x		x	x				x		x	x	x		x	x	x	x	x	x
CAZy annotation					x		x								x			x							x	x	x*		x	x	x	x	x	x
CELLO[(XXIII)]	x															x							x				x		x	x	x	x	x	x

*Only per request. (I) reciprocal; (II) alignment identity; (III) alignment length; (IV) Computational Analysis of gene Family Evolution; (V) for alignment variation detection and mapping; (VI) Burrows–Wheeler Alignment tool; (VII) Multiple Sequence Alignment tool; (VIII) Genome Analysis Toolkit; (IX) Short Oligonucleotide Analysis Package; (X) for repeats in a protein sequence; (XI) Regulatory Sequence Analysis Tool; (XII) An R package; (XIII) transcription factor binding site predictor; (XIV) Secondary Metabolite Unknown Regions Finder; (XV) signal sequence predictor; (XVI) subcellular location prediction tool; (XVII) transmembrane domain prediction tool; (XVIII) for consensus alignment; (XIX) neighbour-joining tree construction; (XX) maximum likelihood tree construction; (XXI) maximum likelihood; (XXII) tree visualization; (XXIII) protein location prediction tool. aa, amino acid.

Methods included in this table are referenced as follows: BLAST (Camacho *et al*., 2009); BLAT (Kent, 2002); BWA (Li and Durbin, 2009); CAFE (De Bie *et al*., 2006); ClustalW/X (Larkin *et al*., 2007); Cosmo (Bembom *et al*., 2007); DIALIGN-TX (Subramanian *et al*., 2010); Emboss (Rice *et al*., 2000); ETE Toolkit (Huerta-Cepas *et al*., 2010); FigTree (Rambaut, 2009); GATK (McKenna *et al*., 2010); Gblocks (Castresana, 2000); InParanoid (O'Brien *et al*., 2005); Kalign (Lassmann and Sonnhammer, 2005); MAFFT (Katoh *et al*., 2002); M-Coffee (Wallace *et al*., 2006); MEGA (Kumar *et al*., 1994); Multi-lagan (Brudno *et al*., 2003); MUMmer (Delcher *et al*., 1999a); MUSCLE (Edgar, 2004); PASA (Haas, 2003); Phylip (Retief, 2000); PhyML (Criscuolo, 2011); PSORT (Horton *et al*., 2007); RaxML (Stamatakis, 2014); Repeat-Masker (http://www.repeatmasker.org/); RepeatScout (Price *et al*., 2005); RSAT (Thomas-Chollier *et al*., 2008); SAMtools (Li *et al*., 2009); SignalP (Emanuelsson *et al*., 2007); SMURF (Khaldi *et al*., 2010); SOAP (Li *et al*., 2008); Sybil (Van De Peer and De Wachter, 1994) (Crabtree *et al*., 2007; Riley *et al*., 2012); Tandem Repeats Finder (Benson, 1999); TMHMM (Emanuelsson *et al*., 2007); TRANSFAC (Wingender *et al*., 2000); Transposon-PSI (Brian Haas, http://transposonpsi.sourceforge.net); TreeCon; TRIBE-MCL (Enright *et al*., 2002); OrthoMCL (Li *et al*., 2003).

should be cautious with using e-values as sole selection criterion, especially when comparing across studies, as it is directly derived from database size and is biased towards long sequence matches and sensitive to data bias. This is particularly challenging in fungi, as there are very large proteins (e.g. polyketide synthases), which may have partial matches to each other, which are longer than some full-length proteins. For this reason, the main motivation for e-value cutoffs should be to reduce the number of false/random hits, and *not* to select true orthologues. This is best achieved by bidirectional hits in tandem with alignment coverage and identity. This also comes with a note of caution. In particular alignment identity is dependent on phylogenetic distance of the species investigated. For this reason, it is not possible to recommend a single set of parameters for BLAST for orthologue detection. Instead we suggest examining the data set and species first, and choose parameters based on the values generated by known orthologues. Typically used parameter settings are described in Table 4.3.

Orthology case studies

A primary application of the identification of putative orthologues is the transfer of annotation information between genomes. In a recent example, the genome of *A. ustus* was sequenced by Pi *et al.* (2015) and the potential protein-coding sequences were found using AUGUSTUS software. These sequences were primarily annotated by homology search against the NCBI non-redundant (nr) protein database (www.ncbi.nlm.nih.gov) using BLASTp with the selection criteria (e-value < $1e^{-03}$, alignment identity > 25%, query alignment coverage > 50%). Their gene ontology annotations were transferred to *A. ustus* when bidirectional best hits between *A. ustus* and *A. nidulans* were fitting the strict parameters (e-value < $1e^{-10}$, alignment identity > 50%, query alignment coverage > 50%) (Pi *et al.*, 2015).

Orthologues have also been used in evolutionary studies. Phylogeny has often been predicted from a single or few conserved genes, but depending on the taxonomic span, the resolution might need to include numerous genes. An approach could be to find all the orthologues shared among the compared species (the core genome) and use these to generate the phylogenetic relation between the species. Investigating shared genes can also give insight into evolutionary development such as horizontal gene transfer and selective pressure.

In a study of pathogenic filamentous fungi, the evolutionary relations between seven *Aspergillus* strains were generated by pairwise comparing genomes using bidirectional best BLASTp hits with an e-value < $1e^{-05}$. From the core genome, 90 genes were selected on the basis of similar lengths and identical number of intron/exon structures. From this analysis Fedorova *et al.* (2008) additionally found genes that potentially were horizontally transferred between *Chaetomium globosum* and *A. fischerianus* (*Neosartorya fischeri*).

Whole-genome comparisons and synteny analysis

Whole genome-to-genome comparison is a highly desired analysis, which includes the study of genetic variations in genes and non-coding regions, chromosomal rearrangements and genetic co-localization (synteny).

Although, the algorithms used today are not ideal as the number of genomes increases (comparison of a large number of genomes is in essence a multidimensional problem) studies with 3–5 whole genome comparisons have not been uncommon (e.g. Fedorova

et al., 2008; Galagan *et al.*, 2005; Rokas *et al.*, 2007). The main issue with increasing the number of genomes is computational time, which increases exponentially with the number of genomes in the comparison, and poses challenges with visualizing the output. Should new algorithms solve this issue, it will allow us to efficiently study more complex dynamics of fungal genomes.

Synteny is commonly addressed using sequence alignments, comparing the localization of genes in one genome with the positions of similar genes in another genome. The alignments are often performed using BLASTp, but other more specialized algorithms as SYBIL (Crabtree *et al.*, 2007; Riley *et al.*, 2012) and MUMmer (Delcher *et al.*, 1999a) have been used (Table 4.3). MUSCLE (Edgar, 2004), MAFFT (Katoh *et al.*, 2002), CLUSTAL X/W (Larkin *et al.*, 2007), and Kalign (Lassmann and Sonnhammer, 2005) are all multiple sequence alignment tools optimized for either long DNA or protein sequences which have been applied in fungal phylogeny research. Additionally, M-Coffee has an extra unique feature where it combines all alignments from other tools generating one consensus alignment (Wallace *et al.*, 2006).

Case examples

Synteny case studies

When the genome of *P. chrysogenum* was released, van den Berg *et al.* (2008) aligned *P. chrysogenum* to *A. nidulans*, *A. niger*, *A. fumigatus*, *and A. oryzae* showing that reshuffling of the chromosomes has occurred after divergence of the *Aspergillus* and *Penicillium* lineages. In particular, they pairwise aligned the *Aspergillus* contigs larger than 100 kb to the 14 largest supercontigs of *P. chrysogenum* using the Promer program of the MUMmer package (van den Berg *et al.*, 2008).

These different cases provide insight to what comparative genomics can be used for. Even if the methods are based on the tools used for single genome analysis, comparing genomes with each other provides the means to study species evolution, gene and chromosomal diversity, physiological traits and novel as well as horizontally transferred genes. Furthermore, comparative genomics provides the opportunity of combining evidence of single genome studies, adding strength to knowledge deduced from the analysis.

Primary metabolism

Primary metabolism can be defined as the set of metabolic reactions directly supporting the propagation of a species by providing precursors for growth and reproduction. Primary metabolism of filamentous fungi is of great interest due to the intriguing phenotype of overproducing large quantities of di- or tri-carboxylic acids derived from intermediates of the citric acid cycle in different species of *Aspergillus*, as well as providing flavour compounds for food production and precursor molecules for production of bioactive compounds from *Penicillium* species. Examples of industrially interesting processes include the production of citric acid by *A. niger* (Karaffa and Kubicek, 2003) or production of itaconic acid by *A. terreus* (Okabe *et al.*, 2009). In order to understand these inter- and intraspecific differences, comparative genomics has been applied as a tool for examining the genomic basis of different phenotypes, e.g. the citric acid-producing strain ATCC 1015 and protein overproducing strain CBS 513.88 of *A. niger* (Andersen *et al.*, 2011). Differences in copy number and

predicted functions can be correlated to productivity and therefore represent an invaluable tool for genetic engineering and target identification. Primary metabolism has been examined on an individual pathway basis in the past, identifying and characterizing the function of individual genes involved in the different pathways.

General comparisons of primary metabolism

Flipphi *et al.* (2009) used the genomes of eight *Aspergillus* species for the comparison of the primary metabolism genes in these organisms. Using this dataset, the authors could identify gene duplications at several steps of primary metabolism indicating the potential of these species to adapt to their individual niches. By comparing multiple genomes the evolution of individual genes can be reconstructed and gene duplication events narrowed down in time.

Genome-scale metabolic models as sources for comparative genomics

Another way of using the wealth of information provided by fungal sequencing efforts is the construction of genome-scale metabolic models. Similar to genome sequencing, where the complete list of genes of a species is enumerated, genome-scale models aim at assembling the complete set of metabolic reactions for a selected species. Due to the unsolved challenge of characterizing metabolic enzymes in a high-throughput manner, understanding differences in primary metabolic phenotype between individual strains or species, relies on external information. In the classical bioinformatics approach, domain and functional predictions are based on sequence motifs and structural properties of the gene products. In addition to this approach, comparative genomics also aims at inferring information of uncharacterized genes from similar characterized genes in a close neighbouring species. Comparative genomics is therefore frequently used during the construction of genome-scale models or in order to fill gaps in known pathways as well as for the construction of less well-characterized organism. There are multiple genome-scale metabolic reconstructions of filamentous fungi available (Table 4.3) and their usage has been reviewed recently (Brandl and Andersen, 2015). Many of these models have been reconstructed using previously published models as templates or source of metabolic reactions (Fig. 4.1). Here, the relatively well-annotated fungal genome of *S. cerevisiae* has served as an initial source of information to generate the first genome-scale models of *Aspergillus* species, which has then propagated to other Aspergilli, and a recent model for *P. chrysogenum.*

This use of comparative genomics propagates biological evidence across species thereby increasing the level of experimental support of the individual models. However, attention should be paid to the possibility of propagating errors between different reconstructions as well as to a potential 'dilution' of the experimental backing. This dilution arises from the fact that gene–function associations are inferred from similarities to a gene in another organism. A cutoff has to be chosen for these comparisons in order to classify gene comparisons as meaningful. By chaining these comparisons the experimental evidence is diluted, as new sequences are no longer compared to the original source but to an inferred gene function itself.

Comparative genomics is not the only source for model generation, as manual reconstructions rely on experimental evidence and bibliomic data, but it is heavily used by algorithms for the automatic generation of genome-scale models (e.g. as described by Pitkänen *et al.*,

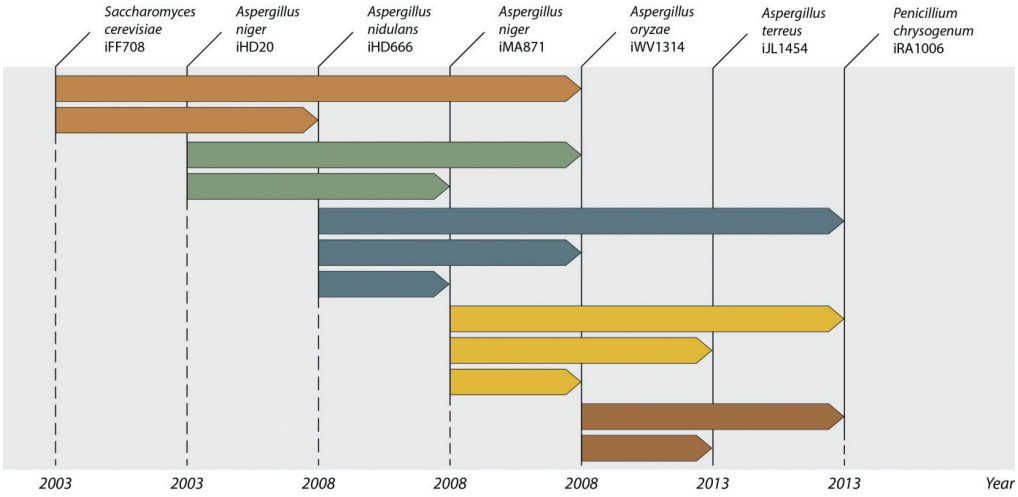


Figure 4.1 Information transfer between different fungal genome-scale models. Propagation of information between the genome-scale models of *S. cerevisiae* (Förster *et al*., 2003), *A. niger* (Andersen *et al*., 2008a; David *et al*., 2003), *A. nidulans* (David *et al*., 2008), *A. oryzae* (Vongsangnak *et al*., 2008), *A. terreus* (Liu *et al*., 2013b) and *P. chrysogenum* (Agren *et al*., 2013).

2014). The combined network from multiple organisms is a strong tool for rapid tentative annotation of primary metabolism in new sequenced organisms.

Secondary metabolism

Comparative genomics is a powerful tool to analyse secondary metabolite producing gene clusters across fungal genomes. In several cases horizontal gene transfer of polyketide synthases (PKS) from bacteria to filamentous fungi has been observed, as well as transfer of gene clusters between fungal species, enabling the fungus to create new compounds by implementation of other genes in the cluster or exchanging cluster members, e.g. in the case of aflatoxin and sterigmatocystin (Osbourn, 2010).

Comparison of gene sequences also led to the identification of new gene clusters, e.g. the aflatrem genes in *A. flavus* have been characterized due to their homology to paxilline from *P. paxilli* (Nicholson *et al.*, 2009), which is an elegant example of how slight modifications in sequence can yield another compound. There are also examples of plant genes, which have been characterized based on fungal sequence information. The cluster coding for helvolic acid synthesis was first identified in *A. fumigatus*, and then cloned into *S. cerevisiae* verifying helvolic acid production. The sequence of the helvolic acid gene cluster was used to determine the biosynthetic genes in plants, shedding light on biosynthetic pathways (Mitsuguchi *et al.*, 2009). Several examples of similar types have been reviewed recently (Sanchez *et al.*, 2012).

Using comparative genomics, one can either search manually (e.g. using sequence alignments) for conserved backbone proteins or their domains in combination with their accessory tailoring genes. Alternatively, one can use predictor tools at the genome level in combination with sequence databases to compare gene clusters throughout organisms.

Table 4.4 presents an overview of such databases and predictors, which are relevant for fungi. A more comprehensive overview also including bacterial tools can be found in Weber *et al.* (2014).

In general, Table 4.4 demonstrates that there are multiple algorithms to choose from, if one wants to comparatively investigate secondary metabolism in fungi. In particular, the predictors can be divided into identifying gene clusters and predicting the specificity of synthases and synthetases in the gene clusters.

Prediction of secondary metabolite gene clusters

In general, there are two main methods applied within the field of predicting secondary metabolite clusters: the Antibiotics and Secondary Metabolite Analysis SHell (AntiSMASH) and the Secondary Metabolite Unique Regions Finder (SMURF) (Khaldi *et al.*, 2010; Weber *et al.*, 2015).

AntiSMASH has recently been upgraded to version 3.0 and optimized to perform faster on large scale projects. The algorithm was built using a profile Hidden Markov Model

Table 4.4 Overview of databases of secondary metabolites, their genes and predictors for fungal secondary metabolite gene clusters

Name	URL	Reference	Organism	Focus[1]
Databases				
Clustermine 360	http://www.clustermine360.ca/	Conway and Boddy, 2013	Fungi, bacteria	Clusters
Norine	http://bioinfo.lifl.fr/norine/	Caboche *et al.*, 2008	Fungi, bacteria	Compound
Novel Antibiotics Database	http://www0.nih.go.jp/~jun/NADB/search.html		Fungi, bacteria	Compound
PubChem	https://pubchem.ncbi.nlm.nih.gov/		Fungi, OTHER	Compound
Prediction of secondary metabolite gene clusters				
antiSMASH3	http://antismash.secondarymetabolites.org/	Weber *et al.*, 2015	Fungi, bacteria	Clusters
SMURF	http://jcvi.org/smurf/index.php	Khaldi *et al.*, 2010	Fungi, bacteria	Clusters
Prediction of substrate specificity				
antiSMASH3	http://antismash.secondarymetabolites.org/	Weber *et al.*, 2015	Fungi, bacteria	PKS, NRPS
MAPSI/ASMPKS	http://gate.smallsoft.co.kr:8008/pks/	Park, 2009	Fungi, bacteria	PKS
NRPSpredictor2	http://nrps.informatik.uni-tuebingen.de/Controller?cmd=SubmitJob	Röttig *et al.*, 2011	Fungi, bacteria	NRPS
LSI-based A-domain function predictor	http://bioserv7.bioinfo.pbf.hr/LSIpredictor/AdomainPrediction.jsp	Baranašić *et al.*, 2014	Bacteria, Fungi	NRPS

[1]PKS, polyketide synthase; NRPS, non-ribosomal peptide synthetase.

(pHMM) on mostly bacterial but also fungal clusters from the NCBI protein database. Once a cluster protein domain (including PKS, nonribosomal peptide synthases (NRPS), terpenes, aminoglycosides, aminocoumarins, indolocarbazoles, lantibiotics, bacteriocins, nucleosides, siderophores or melanin) is found, the program determines accessory genes, by scanning the genome in a 10 kb range and then searches outward by 5, 10 or 20 kb depending on the cluster type. This is highly efficient, but an observed problem with this method is that clusters are predicted to have too many genes included (as commented by Inglis *et al.*, 2013). It rarely underpredicts. A new feature of version 3 is the implementation of ClusterFinder (Cimermancic *et al.*, 2014), which uses a HMM algorithm based on 732 bacterial gene clusters to determine gene clusters of unknown type, filling a gap in the AntiSMASH functionality. However, the training set only consisted of bacterial gene clusters and the performance of this algorithm on fungal sequences still has to be evaluated.

SMURF was launched prior to AntiSMASH, and works in a similar fashion, but only fungal conserved domains were explored to build a HMM algorithm, thus giving it a more specific focus. It also examines a narrower set of cluster domains compared to AntiSMASH, namely dimethyl-allyltransferases including prenyltransferases, hybrids (combinations of PKS and NRPS domains), PKS, PKS-like, NRPS, NRPS-like, terpenecyclases. Once all backbones are found, the algorithm searches for members of 27 secondary metabolite-defining domains within a frame of 20 genes. The user can set intergenic distance between genes in the cluster and the maximum number of secondary metabolite domain negative genes.

In general, AntiSMASH has become sort of a *de facto* standard for secondary metabolite gene cluster predictions, and has developed a user-friendly web interface. This should be applied if one is interested in comparing gene clusters in a given genome. However, given that SMURF is trained on a fungal-specific set, and has a runtime that is significantly below AntiSMASH, we recommend that both algorithms are applied and compared if possible.

Prediction of substrate specificity

With gene clusters identified, one can use a variety of tools to predict the chemicals generated by such clusters (Table 4.4). AntiSMASH (see section on 'Prediction of secondary metabolite gene clusters') includes functions to predict the specificity of both PKSs and NRPSs and is possibly the go-to method at the moment. As an alternative for NRPSs, NRPSpredictor2 (Röttig *et al.*, 2011) uses support vector machines to predict NRPS adenylation domain specificity. The majority of the training set consisted of bacterial gene clusters, while only 16% were of fungal origin. Fortunately, a predictor for fungal gene clusters was added to make investigation of fungal sequences possible.

Supplementing methods for cluster identification

A few methods have been published, which use alternative methods for identifying gene clusters. In particular, (Inglis *et al.*, 2013) used comparative genomics between *A. nidulans*, *A. fumigatus*, *A. niger*, *A. oryzae* to refine the borders of gene clusters returned by antiSMASH and SMURF. Andersen *et al.* (2013) designed an algorithm for the use of transcriptomics data in tandem with genome sequences to determine gene cluster boundaries. A benefit of this method is that it can identify superclusters, i.e. clusters on different chromosomal loci, such as austinol, dehydroaustinol and prenyl xanthones. Additionally, it is not limited by the annotation of the genes, which causes some gene members to be left out in other

algorithms. The MIDDAS-M algorithm (Umemura *et al.*, 2013) offers the implementation of a transcriptomics-driven method using the same principles as the one of (Andersen *et al.*, 2013).

Examples of comparative genomics in secondary metabolism

Until now, comparative studies have shown that only few clusters are conserved in filamentous fungi. Khaldi *et al.* (2010) showed that only five secondary metabolite gene clusters involved in protection against oxidative stress are conserved between the closely related species *A. fumigatus* Af293, *A. clavatus* and *A. fischerianus* (*Neosartorya fischeri*), implying that the other clusters are species-specific. This study showed the power of comparative genomics to reveal common traits of species. However, in this analysis combined with the work by (Lind *et al.*, 2015), one should note that the relatively large diversity in the compared species may underestimate the number of shared clusters. In the future, including new genomes to the analysis could redefine the 'core' clusters and reveal additional common traits.

Chiang *et al.* (2010) employed comparative genomics and protein domain predictions to compare multiple types of PKSs to infer shared traits and modular functions of the proteins, thus identifying shared traits across species.

The study by Inglis *et al.* (2013), mentioned above, has been a first approach to process gene clusters in fungi at larger scale. They identified a total of 261 non-redundant clusters in multiple species using their predicted gene clusters refined by alignment. This is an excellent example of how comparative genomics can complement the quality of automated prediction and can accelerate the research on secondary metabolites in the future.

Secretome analysis and plant polysaccharide degradation

Given the role of filamentous fungi as prolific producers of extracellular enzymes, it is not surprising that a number of studies have employed comparative genomics to investigate the degradation potential of different species. The applications span both the proteases and polysaccharide degrading enzymes.

A general investigation has been performed by (Braaksma *et al.*, 2010). Here, the authors attempted to study the secretome of *A. niger*, and did this by comparing the genome sequences of two *A. niger* strains to *A. oryzae*, *A. fumigatus* and *A. nidulans* to identify all theoretically secreted proteins. This has been combined with a proteomics analysis and allowed a refinement of the identification of secreted proteins from *A. niger*.

Examining proteases has been in a study spanning seven *Aspergillus* species (Ozturkoglu Budak *et al.*, 2014). The authors performed an extensive study of proteases using comparative genomics, comparative proteomics, and protease assays. The study confirmed that the aspergilli examined have protease capacities of similar magnitude, and that the different species have divergent profiles of proteases dependent on the growth medium, thus suggesting specialized life styles.

Polysaccharide degradation potential is a trait of interest in most species, and has been uncovered in many of the recent genome publications (Ballester *et al.*, 2015; Marcet-Houben *et al.*, 2012; Pel *et al.*, 2007). For a more specialized investigation, Coutinho *et al.* (2009) examined the polysaccharide degradation potential of three *Aspergillus* species in detail, using a combination of CAZy predictions (Cantarel *et al.*, 2009; Lombard *et al.*, 2014), comparative genomics, and growth assays on polysaccharides. Furthermore, the

authors included a comparison to *Podospora anserina*. The main discovery in this study was the finding that the fungi had relatively similar degradation potential, but the actual degradation profiles of the fungi were different, possibly due to differential regulation and natural ecological niches.

Other comparative 'omics'

In the context of comparative genetics, it can be quite informative to also examine other types of comparative omics and examine what can be learned from these approaches. In this case, we see comparative omics as performing the same type of 'omics' analysis in multiple species and comparing the results using an identification of orthologues between the species. As the genome is used for orthologue determination (see 'Identification of orthologues across species', above), it can be seen as an additional dimension added to the comparative genomics data by an experimental analysis. Some studies of this type have been performed, in particular comparative transcriptomics and proteomics.

Comparative transcriptomics

Comparative transcriptomics is particularly interesting, when one wishes to examine and identify conserved transcriptional regulation across species. To our knowledge the first example of this in *Aspergillus* and/or *Penicillium* was by Andersen *et al.* (2008), where the regulatory response to growing on L-xylose relative to D-glucose was examined. The authors carried out DNA microarray experiments for *A. niger*, *A. nidulans* and *A. oryzae*, identified a set of genes with the same transcriptional response in all three examined species, and a common regulatory motif for binding of the xylanolytic transcriptional regulator (XlnR).

Salazar *et al.* (2009) performed a similar study using glucose and glycerol as the carbon sources, which gave interesting findings on the fungal response to osmotic stress. It was seen that the *Aspergillus* response to glycerol-induced osmotic stress has conserved elements and regulatory motifs to similar systems in *S. cerevisiae* and humans.

The study by Coutinho *et al.* (2009), which we discussed in the section on 'Secretome analysis and plant polysaccharide degradation', also employs comparative transcriptomics based on the data set of Andersen *et al.* (2008b) to study the secretory response and regulation, concluding a differential regulation of secreted polysaccharide-degrading enzymes.

The responses to hypoxia in *A. oryzae* and *A. nidulans* have been studied by Terabayashi *et al.* (2012) by comparing DNA microarray experiments of *A. oryzae* with available data for *A. nidulans.* Cultures were performed under hypoxic conditions and identified both conserved and species-specific responses to hypoxia.

Comparative proteomics

In this context, it is important to define that we here see 'comparative proteomics' as the comparison of transcriptomic results from multiple species or diverse strains, and not, as is also seen in the literature, the comparison of proteomics from multiple conditions [e.g. Stoll *et al.* (2014) or other notable papers as also reviewed by Kniemeyer (2011)].

Within our definition of comparative proteomics, an interesting aspect is, that this approach allows identification of conserved responses as well as unique traits across multiple species.

One such example in the area of immunogenicity, is the work of Benndorf *et al.* (2008), who studied allergens from spores of *A. versicolor*, *A. fumigatus*, and *P. expansum* by performing 2D gels in combination with immunoblotting to identify allergenic proteins in individual and multiple species.

A very comprehensive study of secreted proteases in seven different *Aspergillus* species has been conducted (Ozturkoglu Budak *et al.*, 2014). Here, the authors mine the genomes for putative proteases, employ comparative genomics to identify shared and unique proteases, and followed up by conducting proteome profiling on the secreted proteins from all seven species. Results were further validated with enzymatic protease assays. The analysis identified highly conserved proteins across the species, but with very different expression profiles even on the same media. An important hypothesis was that the species may have different physiologies at the same time point although cultivated in the same manner; similar to what was concluded for the polysaccharide-degrading enzymes by Coutinho *et al.* (2009).

In general, both comparative transcriptomics and proteomics, while expensive to conduct, have the advantage of letting us extrapolate what is known in one species across multiple species, thus convincingly allowing the researchers to identify whether a specific trait is unique for a species or a general biological phenomenon. Such studies have interesting potentials for understanding biology at a more general level.

Future perspectives – what can we learn from bacteria?

If we are to guess at what the future of fungal comparative genetics will bring, it is tempting to look at the kingdom of bacteria. Genome sequencing started in bacteria 10 years before Aspergilli and Penicillia with the first sequencing of the prokaryote *Haemophilus influenzae* in 1995. As a consequence of this head start, and the advantage of the lower price of sequencing the smaller bacterial genomes, the number of bacterial genomes now number in the tens of thousands (Land *et al.*, 2014), and some individual species have more than 2000 isolates sequenced (Cook and Ussery, 2013). Furthermore, by the time the first fungal genome was published in 2005, sequence databases, online analysis tools, gene calling pipelines and functional annotation models had been created and optimized for bacteria and were applied to fungal genomes. In some aspects, the technology of analysing bacterial genomes is both the future and the past of fungal comparative genomics.

Data sharing, maintenance and standards

An area, where the experience from bacterial genomics has been a great advantage, is in the construction and use of sequence databases. Through resources like GenBank, SWISS-PROT and RefSeq, experiences and ideas have been gained and used to create systems with more optimized downloads, searches, data submission and community involvement. Optimal design of data sharing resources has continually led to the introduction of new setups in resources like GenBank (Wheeler *et al.*, 2013). The fungal community has the opportunity to take advantage of these experiences and create databases that suit the needs of the research community. Some of this work has already been addressed in the structure and integration of FungiDB (Stajich *et al.*, 2012) and EuPathDB (Aurrecoechea *et al.*, 2013) as discussed above (section on 'Available genome databases with comparative genomics capabilities').

A key component in setting up data resources is data standards, for downloads. Although different data formats are widely used (FASTA, TAB, XML) the specifications on what information to include in these formats are minimal. Problems include naming of organisms, identification numbers of organisms and genes, relations between gene and protein sequences as well as metadata like literature references and growth conditions. All of these problems make it difficult to relate data resources to one another and utilize the full potential of these efforts. Comparative genomics relies heavily on the availability and quality of data without which comparisons become biased and less informative (Schnoes *et al.*, 2013).

The genetic data from sequencing of fungal genomes will increase rapidly in the coming years and efforts should be made to establish community standards for data, metadata, data submission and download to ensure the future use and reproducibility of *in silico* analyses. Efforts should also be made to encourage incorporation of experimental knowledge with sequence data to gain maximum information and understanding from years of experimental knowledge and future sequencing efforts.

The increased number of data and the pace at which it is being generated, presents a technical problem for the field of bioinformatics. In recent years the rate at which sequencing data has grown exceeds the pace of Moore's law (a doubling every 2 years). This puts a new pressure on the development of methods that can deal with large amounts of data without relying on larger computers. The issue of storage must also be addressed, with funding being granted to maintain online resources and publishing houses insisting on data submissions for research that does *in silico* analyses.

Comparative genomics-driven insights

Comparative genomics has been used in a wide range of bacterial research areas and some applications have yet to be widely studied in fungal research. One such topic is the genetic variation of species, genus and family. Some attention has been given to the core genome concept, identifying genes found in all strains of a specific set (Fedorova *et al.*, 2008; Galagan *et al.*, 2005). Genes found in specific strain subsets can also be used to connect possible phenotypes to genotypes, aiding experimental design (Cheeseman *et al.*, 2014; Coutinho *et al.*, 2009). However, these studies have been limited by the number of available genomes and general genetic variance of fungi is yet to be investigated. As more data becomes available along with well-documented metadata, it will be possible to use comparative genomics to identify genes involved in specific traits such as pathogenicity or identify sets of strain-specific genes. Comparative genomics can also aid in construction of evolutionary models by including whole genome patterns, adding deeper resolution to the known taxonomy (Rokas *et al.*, 2007; Snipen and Ussery, 2010), especially in the light of the expansion in number of fungal genome sequences discussed in the section on 'Current status of genomics'.

The role of fungi in biotechnology has long made them target for optimization of production. One comparative genomics-driven method of optimizing this is modelling of primary metabolism, another feature which has primarily been successful in bacteria and yeast (Heavner *et al.*, 2013; Monk *et al.*, 2013; Orth *et al.*, 2011). Genome-scale metabolic models will facilitate metabolic engineering by offering *in silico* predicted targets before experimental testing (Brandl and Andersen, 2015). Comparative genomics can greatly increase the evidence for these targets by looking for patterns across multiple genome sequences.

The general concept of a model, a mathematical description of a pattern in data, can in theory be used to describe almost any kind of pattern assuming that enough example data is

provided. Creating models requires training data that represents enough variation to generalized the pattern (Ansari *et al.*, 2004). Sequence models can for example be used to predict secreted proteins (Braaksma *et al.*, 2010) or identify carbohydrate-active enzymes (CAZy) (Cantarel *et al.*, 2009).

Genome sequencing and annotation

A fungal research area that has extensively used comparative genomics is the field of secondary metabolism. High-throughput *in silico* screening of fungal genome sequences has revealed a large number of possible gene clusters involved in secondary metabolism, opening up for targeted experimental studies and discovery of new compounds for industrial production. The topics addressed including identification of novel clusters (Chiang *et al.*, 2009; Hansen *et al.*, 2011; Inglis *et al.*, 2013), locating known clusters in new genomes, assigning functions to genes in clusters (Szewczyk *et al.*, 2008) and locating silent gene clusters (Bergmann *et al.*, 2007). The use of *in silico* methods within this field has developed rapidly, including machine learning approaches as well as manually curated data collections (Weber, 2014).

Secondary metabolism has also been studied extensively in bacteria, exploiting the relative ease of sequencing many of these species for genome mining (Duncan *et al.*, 2015). However, despite this wealth of data and studies, there are still problems in the field, in particular with gene finding. Problems include inconsistencies due to different sequencing technologies or quality as well as use of different gene callers. Some of the most common problems involve over-annotation – prediction of genes that are not actually expressed. There is currently no clear way of determining *in silico* if a gene is truly a gene, experiments can reveal if a gene is expressed under some conditions but testing that a gene is not expressed under any conditions is not feasible on any scale. Comparative genomics has been used to address some of the annotation problems including assembly of genomes based on other genomes and gap filling (Delcher *et al.*, 1999a) and annotation, gene finding and functions (Inglis *et al.*, 2013).

The scale of over- or under-annotation becomes increasingly obvious as more genomes are sequenced. In one study, 4.94 million genes were identified *in silico* in 1474 prokaryotic organisms and 7% of these were not found in the published annotations (Wood *et al.*, 2012). The study showed that 60% of these genes were genes of length between 110 and 300 amino acids, thus suggesting that short genes in general are underpredicted. Another aspect of gene finding is the setting of parameters used for different prediction programs. Although some settings are widely used, actual benchmarking of different settings and the effect of these should be performed and published for the general community as these can have a major effect on the analysis outcome (Mancheron *et al.*, 2011).

All of these topics offer types of information and data ideal for modelling and machine learning. Data driven models have been extensively used in bacterial research to describe and identify functional protein domains such as PFAM or PANTHER (Finn *et al.*, 2014; Mi *et al.*, 2005) or creating gene calling models such as GLIMMER, SnowyOwl or Prodigal (Delcher *et al.*, 1999b; Hyatt *et al.*, 2010; Reid *et al.*, 2014). The success of these methods relies heavily on available training data, connections to experimental data and manual curation. Although these tools can be used for fungal genomes and gene sequences, the training data is currently biased towards bacteria due to the larger number of available data (e.g. AntiSMASH). In order for fungal annotation to improve, more fungal data should be included

in the model training as well as manual curation of literature and sequence models. Creating new tools specifically for fungi is also an option, as in the SnowyOwl gene calling algorithm (Reid *et al.*, 2014).

We strongly believe that the post-genome future of fungal genomes will be highly driven by the application of comparative genomics approaches. The future of fungal comparative genomics promises to produce more and better data, addressing broader research fields and integrating into more projects. Establishing good practices on how this integration is created, how it should be published, documented and shared will be vital for its success and contribution to research. Making use of the experiences from other fields should shape the way the fungal community deals with these issues and affect the way grants are funded, publications are written and projects are planned.

References

Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I., and Nielsen, J. (2013). The RAVEN Toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. PLoS Comput. Biol. *9*.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Andersen, M., Nielsen, M., and Nielsen, J. (2008a). Metabolic model integration of the bibliome, genome, metabolome and reactome of Aspergillus niger. Mol. Syst. Biol. *4*, 178.

Andersen, M.R., Vongsangnak, V., Panagiotou, G., Salazar, M.P., Lehmann, L., and Nielsen, J. (2008b). A trispecies *Aspergillus* microarray: comparative transcriptomics of three Aspergillus species. Proc. Natl. Acad. Sci. U.S.A. *105*, 4387–4392.

Andersen, M.R., Salazar, M.P., Schaap, P.J., van de Vondervoort, P.J.I., Culley, D., Thykaer, J., Frisvad, J.C., Nielsen, K.F., Albang, R., Albermann, K., *et al.* (2011). Comparative genomics of citric-acid-producing *Aspergillus niger* ATCC 1015 versus enzyme-producing CBS 513.88. Genome Res. *21*, 885–897.

Andersen, M.R., Nielsen, J.B., Klitgaard, A., Petersen, L.M., Zachariasen, M., Hansen, T.J., Blicher, L.H., Gotfredsen, C.H., Larsen, T.O., Nielsen, K.F., *et al.* (2013). Accurate prediction of secondary metabolite gene clusters in filamentous fungi. Proc. Natl. Acad. Sci. U.S.A. *110*, E99–E107.

Ansari, M.Z., Yadav, G., Gokhale, R.S., and Mohanty, D. (2004). NRPS-PKS: A knowledge-based resource for analysis of NRPS-PKS megasynthases. Nucleic Acids Res. *32*.

Arnaud, M.B., Chibucos, M.C., Costanzo, M.C., Crabtree, J., Inglis, D.O., Lotia, A., Orvis, J., Shah, P., Skrzypek, M.S., Binkley, G., *et al.* (2010). The Aspergillus Genome Database, a curated comparative genomics resource for gene, protein and sequence information for the Aspergillus research community. Nucleic Acids Res. *38*, D420–D427.

Aurrecoechea, C., Barreto, A., Brestelli, J., Brunk, B.P., Cade, S., Doherty, R., Fischer, S., Gajria, B., Gao, X., Gingle, A., *et al.* (2013). EuPathDB: The eukaryotic pathogen database. Nucleic Acids Res. *41*, 684–691.

Ballester, A., Marcet-houben, M., Levin, E., Sela, N., Selma-lázaro, C., Carmona, L., Wisniewski, M., Droby, S., González-candelas, L., and Gabaldón, T. (2015). Genome, transcriptome, and functional analyses of *Penicillium expansum* provide new insights into secondary metabolism and pathogenicity. Mol. Plant-Microbe Interact. *28*, 232–248.

Baranašić, D., Zucko, J., Diminic, J., Gacesa, R., Long, P.F., Cullum, J., Hranueli, D., and Starcevic, A. (2014). Predicting substrate specificity of adenylation domains of nonribosomal peptide synthetases and other protein properties by latent semantic indexing. J. Ind. Microbiol. Biotechnol. *41*, 461–467.

Bembom, O., Keles, S., and van der Laan, M.J. (2007). Supervised detection of conserved motifs in DNA sequences with cosmo. Stat. Appl. Genet. Mol. Biol. *6*, Article 8.

Benndorf, D., Müller, a., Bock, K., Manuwald, O., Herbarth, O., and Von Bergen, M. (2008). Identification of spore allergens from the indoor mould Aspergillus versicolor. Allergy Eur. J. Allergy Clin. Immunol. *63*, 454–460.

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580.

van den Berg, M.A., Albang, R., Albermann, K., Badger, J.H., Daran, J.-M., Driessen, A.J.M., Garcia-Estrada, C., Fedorova, N.D., Harris, D.M., Heijne, W.H.M., *et al.* (2008). Genome sequencing and analysis of the filamentous fungus *Penicillium chrysogenum*. Nat. Biotechnol. *26*, 1161–1168.

Bergmann, S., Schümann, J., Scherlach, K., Lange, C., Brakhage, A. a, and Hertweck, C. (2007). Genomics-driven discovery of PKS–NRPS hybrid metabolites from *Aspergillus nidulans*. Nat. Chem. Biol. *3*, 213–217.

Berry, D., Cox, M.P., and Scott, B. (2015). Draft genome sequence of the filamentous fungus *Penicillium paxilli* (ATCC 26601). Genome Announc. *3*, e00071–15.

De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics *22*, 1269–1271.

Bok, J.W., Hoffmeister, D., Maggio-Hall, L.A., Murillo, R., Glasner, J.D., and Keller, N.P. (2006). Genomic mining for Aspergillus natural products. Chem. Biol. *13*, 31–37.

Braaksma, M., Martens-Uzunova, E.S., Punt, P.J., and Schaap, P.J. (2010). An inventory of the *Aspergillus niger* secretome by combining *in silico* predictions with shotgun proteomics data. BMC Genomics *11*, 584.

Brandl, J., and Andersen, M.R. (2015). Current state of genome-scale modeling in filamentous fungi. Biotechnol. Lett. *37*, 1131–1139

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., and Batzoglou, S. (2003). LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. *13*, 721–731.

Caboche, S., Pupin, M., Leclère, V., Fontaine, A., Jacques, P., and Kucherov, G. (2008). NORINE: a database of nonribosomal peptides. Nucleic Acids Res. *36*, D326–D331.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res. *37*, D233–D238.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. *17*, 540–552.

Cerqueira, G.C., Arnaud, M.B., Inglis, D.O., Skrzypek, M.S., Binkley, G., Simison, M., Miyasato, S.R., Binkley, J., Orvis, J., Shah, P., *et al.* (2014). The Aspergillus Genome Database: Multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. Nucleic Acids Res. *42*, 705–710.

Cheeseman, K., Ropars, J., Renault, P., Dupont, J., Gouzy, J., Branca, A., Abraham, A.-L., Ceppi, M., Conseiller, E., Debuchy, R., *et al.* (2014). Multiple recent horizontal transfers of a large genomic region in cheese making fungi. Nat. Commun. *5*, 2876.

Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., *et al.* (2012). Saccharomyces Genome Database: The genomics resource of budding yeast. Nucleic Acids Res. *40*.

Chiang, Y.-M., Szewczyk, E., Davidson, A.D., Keller, N., Oakley, B.R., and Wang, C.C.C. (2009). A gene cluster containing two fungal polyketide synthases encodes the biosynthetic pathway for a polyketide, asperfuranone, in *Aspergillus nidulans*. J. Am. Chem. Soc. *131*, 2965–2970.

Chiang, Y.-M., Oakley, B.R., Keller, N.P., and Wang, C.C.C. (2010). Unraveling polyketide synthesis in members of the genus Aspergillus. Appl. Microbiol. Biotechnol. *86*, 1719–1736.

Cimermancic, P., Medema, M.H., Claesen, J., Kurita, K., Wieland Brown, L.C., Mavrommatis, K., Pati, A., Godfrey, P.A., Koehrsen, M., Clardy, J., *et al.* (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell *158*, 412–421.

Conway, K.R., and Boddy, C.N. (2013). ClusterMine360: A database of microbial PKS/NRPS biosynthesis. Nucleic Acids Res. *41*, 402–407.

Cook, H., and Ussery, D.W. (2013). Sigma factors in a thousand *E. coli* genomes. Environ. Microbiol. *15*, 3121–3129.

Coutinho, P.M., Andersen, M.R., Kolenova, K., VanKuyk, P.A., Benoit, I., Gruben, B.S., Trejo-Aguilar, B., Visser, H., van Solingen, P., Pakula, T., *et al.* (2009). Post-genomic insights into the plant polysaccharide degradation potential of *Aspergillus nidulans* and comparison to *Aspergillus niger* and *Aspergillus oryzae*. Fungal Genet. Biol. *46* (Suppl. 1), S161–S169.

Crabtree, J., Angiuoli, S.V., Wortman, J.R., and White, O.R. (2007). Sybil: methods and software for multiple genome comparison and visualization. Methods Mol. Biol. *408*, 93–108.

Criscuolo, A. (2011). MorePhyML: Improving the phylogenetic tree space exploration with PhyML 3. Mol. Phylogenet. Evol. *61*, 944–948.
Dalquen, D.A., and Dessimoz, C. (2013). Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. Genome Biol. Evol. *5*, 1800–1806.
David, H., Akesson, M., and Nielsen, J. (2003). Reconstruction of the central carbon metabolism of *Aspergillus niger*. Eur. J. Biochem. *270*, 4243–4253.
David, H., Ozçelik, I.S., Hofmann, G., and Nielsen, J. (2008). Analysis of *Aspergillus nidulans* metabolism at the genome-scale. BMC Genomics *9*, 163.
Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O., and Salzberg, S.L. (1999a). Alignment of whole genomes. Nucleic Acids Res. *27*, 2369–2376.
Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. (1999b). Improved microbial gene identification with GLIMMER. Nucleic Acids Res. *27*, 4636–4641.
Duncan, K.R., Crüsemann, M., Lechner, A., Sarkar, A., Li, J., Ziemert, N., Wang, M., Bandeira, N., Moore, B.S., Dorrestein, P.C., *et al.* (2015). Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from Salinispora species. Chem. Biol.
Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.
Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. Nat. Protoc. *2*, 953–971.
Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. *30*, 1575–1584.
Fedorova, N.D., Khaldi, N., Joardar, V.S., Maiti, R., Amedeo, P., Anderson, M.J., Crabtree, J., Silva, J.C., Badger, J.H., Albarraq, A., *et al.* (2008). Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. PLoS Genet. *4*, e1000046.
Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., *et al.* (2014). Pfam: The protein families database. Nucleic Acids Res. *42*.
Flipphi, M., Sun, J., Robellet, X., Karaffa, L., Fekete, E., Zeng, A.-P., and Kubiecek, C.P. (2009). Biodiversity and evolution of primary carbon metabolism in *Aspergillus nidulans* and other *Aspergillus* spp. Fungal Genet. Biol. 46(Suppl. 1), S19–S44.
Förster, J., Famili, I., Fu, P., Palsson, B., and Nielsen, J. (2003). Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. Genome Res. *13*, 244–253.
Futagami, T., Mori, K., Yamashita, A., Wada, S., Kajiwara, Y., Takashita, H., Omori, T., Takegawa, K., Tashiro, K., Kuhara, S., *et al.* (2011). Genome sequence of the white koji mold Aspergillus kawachii IFO 4308, used for brewing the Japanese distilled spirit shochu. Eukaryot. Cell *10*, 1586–1587.
Galagan, J.E., Calvo, S.E., Cuomo, C., Ma, L.-J., Wortman, J.R., Batzoglou, S., Lee, S.-I., Baştürkmen, M., Spevak, C.C., Clutterbuck, J., *et al.* (2005). Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. Nature *438*, 1105–1115.
Gilsenan, J.M., Cooley, J., and Bowyer, P. (2012). CADRE: The Central Aspergillus Data REpository 2012. Nucleic Acids Res. *40*, 660–666.
Grigoriev, I.V., Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D., Kuo, A., Minovitsky, S., Nikitin, R., Ohm, R.A., *et al.* (2012). The genome portal of the Department of Energy Joint Genome Institute. Nucleic Acids Res *40*, D26–D32.
Grigoriev, I.V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., Riley, R., Salamov, A., Zhao, X., Korzeniewski, F., *et al.* (2014). MycoCosm portal: Gearing up for 1000 fungal genomes. Nucleic Acids Res. *42*, 699–704.
Haas, B.J. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. *31*, 5654–5666.
Hansen, B.G., Genee, H.J., Kaas, C.S., Nielsen, J.B., Regueira, T.B., Mortensen, U.H., Frisvad, J.C., and Patil, K.R. (2011). A new class of IMP dehydrogenase with a role in self-resistance of mycophenolic acid producing fungi. BMC Microbiol. *11*, 202.
Heavner, B.D., Smallbone, K., Price, N.D., and Walker, L.P. (2013). Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance. Database (Oxford) *2013*, bat059.
Horton, P., Park, K.J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J., and Nakai, K. (2007). WoLF PSORT: Protein localization predictor. Nucleic Acids Res. *35*.
Huerta-Cepas, J., Dopazo, J., and Gabaldón, T. (2010). ETE: a python Environment for Tree Exploration. BMC Bioinformatics *11*, 24.

Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119.
Inglis, D.O., Binkley, J., Skrzypek, M.S., Arnaud, M.B., Cerqueira, G.C., Shah, P., Wymore, F., Wortman, J.R., and Sherlock, G. (2013). Comprehensive annotation of secondary metabolite biosynthetic genes and gene clusters of *Aspergillus nidulans*, *A. fumigatus*, *A. niger* and *A. oryzae*. BMC Microbiol *13*, 91.
Karaffa, L., and Kubicek, C.P. (2003). Aspergillus niger citric acid accumulation: do we understand this well working black box? Appl. Microbiol. Biotechnol. *61*, 189–196.
Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. *30*, 3059–3066.
Kent, W.J. (2002). BLAT – The BLAST-Like Alignment Tool. Genome Res. *12*, 656–664.
Khaldi, N., Seifuddin, F.T., Turner, G., Haft, D., Nierman, W.C., Wolfe, K.H., and Fedorova, N.D. (2010). SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet. Biol. *47*, 736–741.
Kis-Papo, T., Weig, A.R., Riley, R., Peršoh, D., Salamov, A., Sun, H., Lipzen, A., Wasser, S.P., Rambold, G., Grigoriev, I.V., *et al.* (2014). Genomic adaptations of the halophilic Dead Sea filamentous fungus Eurotium rubrum. Nat. Commun. *5*, 3745.
Kniemeyer, O. (2011). Proteomics of eukaryotic microorganisms: The medically and biotechnologically important fungal genus *Aspergillus*. Proteomics *11*, 3232–3243.
Kumar, S., Tamura, K., and Nei, M. (1994). MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. Comput. Appl. Biosci. *10*, 189–191.
Land, M.L., Hyatt, D., Jun, S.-R., Kora, G.H., Hauser, L.J., Lukjancenko, O., and Ussery, D.W. (2014). Quality scores for 32,000 genomes. Stand. Genomic Sci. *9*, 20.
Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., *et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947–2948.
Lassmann, T., and Sonnhammer, E.L.L. (2005). Kalign – an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics *6*, 298.
Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
Li, B., Zong, Y., Du, Z., Chen, Y., Zhang, Z., Qin, G., Zhao, W., and Tian, S. (2015). Genomic characterization reveals insights into patulin biosynthesis and pathogenicity in *Penicillium* species. Mol. Plant Microbe Interact.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
Li, L., Stoeckert, C.J., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. *13*, 2178–2189.
Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008). SOAP: Short oligonucleotide alignment program. Bioinformatics *24*, 713–714.
Lind, A.L., Wisecaver, J.H., Smith, T.D., Feng, X., Calvo, A.M., and Rokas, A. (2015). Examining the evolution of the regulatory circuit controlling secondary metabolism and development in the fungal genus *Aspergillus*. PLOS Genet. *11*, e1005096.
Linz, J.E., Wee, J., and Roze, L.V. (2014). *Aspergillus parasiticus* SU-1 genome sequence, predicted chromosome structure, and comparative gene expression under aflatoxin-inducing conditions: evidence that differential expression contributes to species phenotype. Eukaryot. Cell *13*, 1113–1123.
Liu, G., Zhang, L., Wei, X., Zou, G., Qin, Y., Ma, L., Li, J., Zheng, H., Wang, S., Wang, C., *et al.* (2013a). Genomic and secretomic analyses reveal unique features of the lignocellulolytic enzyme system of *Penicillium decumbens*. PLoS One *8*.
Liu, J., Gao, Q., Xu, N., and Liu, L. (2013b). Genome-scale reconstruction and *in silico* analysis of *Aspergillus terreus* metabolism. Mol. Biosyst. *9*, 1939–1948.
Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. *42*, 490–495.
Machida, M., Asai, K., Sano, M., Tanaka, T., Kumagai, T., Terai, G., Kusumoto, K.-I., Arima, T., Akita, O., Kashiwagi, Y., *et al.* (2005). Genome sequencing and analysis of *Aspergillus oryzae*. Nature *438*, 1157–1161.
Mancheron, A., Uricaru, R., and Rivals, E. (2011). An alternative approach to multiple genome comparison. Nucleic Acids Res. 1–11.
Marcet-Houben, M., Ballester, A.-R., de la Fuente, B., Harries, E., Marcos, J.F., González-Candelas, L., and Gabaldón, T. (2012). Genome sequence of the necrotrophic fungus Penicillium digitatum, the main postharvest pathogen of citrus. BMC Genomics *13*, 646.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.* (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M.J., *et al.* (2005). The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res. *33*.

Mitsuguchi, H., Seshime, Y., Fujii, I., Shibuya, M., Ebizuka, Y., and Kushiro, T. (2009). Biosynthesis of steroidal antibiotic fusidanes: Functional analysis of oxidosqualene cyclase and subsequent tailoring enzymes from *Aspergillus fumigatus*. J. Am. Chem. Soc. *131*, 6402–6411.

Monk, J.M., Charusanti, P., Aziz, R.K., Lerman, J.A., Premyodhin, N., Orth, J.D., Feist, A.M., and Palsson, B.Ø. (2013). Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. Proc. Natl. Acad. Sci. U.S.A. *110*, 20338–20343.

Nicholson, M.J., Koulman, A., Monahan, B.J., Pritchard, B.L., Payne, G.A., and Scott, B. (2009). Identification of two aflatrem biosynthesis gene loci in *Aspergillus flavus* and metabolic engineering of *Penicillium paxilli* to elucidate their function. Appl. Environ. Microbiol. *75*, 7469–7481.

Nierman, W.C., Pain, A., Anderson, M.J., Wortman, J.R., Kim, H.S., Arroyo, J., Berriman, M., Abe, K., Archer, D.B., Bermejo, C., *et al.* (2005). Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. Nature *438*, 1151–1156.

Nierman, W.C., Yu, J., Fedorova-Abrams, N.D., Losada, L., Cleveland, T.E., Bhatnagar, D., Bennett, J.W., Dean, R., and Payne, G.A. (2015a). Genome sequence of *Aspergillus flavus* NRRL 3357, a strain that causes aflatoxin contamination of food and feed. Genome Announc. *3*, e00168–15.

Nierman, W.C., Fedorova-Abrams, N.D., and Andrianopoulos, A. (2015b). Genome sequence of the AIDS-associated pathogen *Penicillium marneffei* (ATCC18224) and its near taxonomic relative *Talaromyces stipitatus* (ATCC10500). Genome Announc. *3*, e01559–14.

O'Brien, K.P., Remm, M., and Sonnhammer, E.L.L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. *33*, D476–D480.

Okabe, M., Lies, D., Kanamasa, S., and Park, E.Y. (2009). Biotechnological production of itaconic acid and its biosynthesis in Aspergillus terreus. Appl. Microbiol. Biotechnol. *84*, 597–606.

Orth, J.D., Conrad, T.M., Na, J., Lerman, J.A., Nam, H., Feist, A.M., and Palsson, B.Ø. (2011). A comprehensive genome-scale reconstruction of *Escherichia coli* Metabolism. Mol. Syst. Biol. *7*, 535.

Osbourn, A. (2010). Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. Trends Genet. *26*, 449–457.

Ozturkoglu Budak, S., Zhou, M., Brouwer, C., Wiebenga, A., Benoit, I., Di Falco, M., Tsang, A., and de Vries, R.P. (2014). A genomic survey of proteases in aspergilli. BMC Genomics *15*, 523.

Park, K. (2009). Development of an analysis program of type I polyketide synthase gene clusters using homology search and profile hidden Markov model. J. Microbiol. Biotechnol. *19*, 140–146.

Van De Peer, Y., and De Wachter, R. (1994). Treecon for windows: A software package for the construction and drawing of evolutionary trees for the microsoft windows environment. Bioinformatics *10*, 569–570.

Pel, H.J., de Winde, J.H., Archer, D.B., Dyer, P.S., Hofmann, G., Schaap, P.J., Turner, G., de Vries, R.P., Albang, R., Albermann, K., *et al.* (2007). Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. Nat. Biotechnol. *25*, 221–231.

Pi, B., Yu, D., Dai, F., Song, X., Zhu, C., Li, H., and Yu, Y. (2015). A genomics based discovery of secondary metabolite biosynthetic gene clusters in *Aspergillus ustus*. PLoS One *10*, e0116089.

Pitkänen, E., Jouhten, P., Hou, J., Syed, M.F., Blomberg, P., Kludas, J., Oja, M., Holm, L., Penttilä, M., Rousu, J., *et al.* (2014). Comparative genome-scale reconstruction of gapless metabolic networks for present and ancestral species. PLoS Comput. Biol. *10*, e1003465.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). *De novo* identification of repeat families in large genomes. Bioinformatics *21*.

Rambaut, A. (2009). FigTree, a graphical viewer of phylogenetic trees. Inst. Evol. Biol. Univ. Edinburgh.

Reid, I., O'Toole, N., Zabaneh, O., Nourzadeh, R., Dahdouli, M., Abdellateef, M., Gordon, P.M.K., Soh, J., Butler, G., Sensen, C.W., *et al.* (2014). SnowyOwl: accurate prediction of fungal genes by using RNA-Seq and homology information to select among *ab initio* models. BMC Bioinformatics *15*, 229.

Retief, J.D. (2000). Phylogenetic analysis using PHYLIP. Methods Mol. Biol. *132*, 243–258.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. *16*, 276–277.

Riley, D.R., Angiuoli, S.V., Crabtree, J., Dunning Hotopp, J.C., and Tettelin, H. (2012). Using Sybil for interactive comparative genomics of microbes on the web. Bioinformatics *28*, 160–166.

Rokas, A., Payne, G., Fedorova, N.D., Baker, S.E., Machida, M., Yu, J., Georgianna, D.R., Dean, R.A., Bhatnagar, D., Cleveland, T.E., *et al.* (2007). What can comparative genomics tell us about species concepts in the genus *Aspergillus*? Stud. Mycol. *59*, 11–17.

Röttig, M., Medema, M.H., Blin, K., Weber, T., Rausch, C., and Kohlbacher, O. (2011). NRPSpredictor2 – A web server for predicting NRPS adenylation domain specificity. Nucleic Acids Res. *39*.

Salazar, M., Vongsangnak, W., Panagiotou, G., Andersen, M., and Nielsen, J. (2009). Uncovering transcriptional regulation of glycerol metabolism in Aspergilli through genome-wide gene expression data analysis. Mol. Genet. Genomics *282*, 571–586.

Sanchez, J.F., Somoza, A.D., Keller, N.P., and Wang, C.C.C. (2012). Advances in *Aspergillus* secondary metabolite research in the post-genomic era. Nat. Prod. Rep. *29*, 351–371.

Schnoes, A.M., Ream, D.C., Thorman, A.W., Babbitt, P.C., and Friedberg, I. (2013). Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. PLoS Comput. Biol. *9*, e1003063.

Snipen, L., and Ussery, D.W. (2010). Standard operating procedure for computing pangenome trees. Stand. Genomic Sci. *2*, 135–141.

Stajich, J.E., Harris, T., Brunk, B.P., Brestelli, J., Fischer, S., Harb, O.S., Kissinger, J.C., Li, W., Nayak, V., Pinney, D.F., *et al.* (2012). FungiDB: An integrated functional genomics database for fungi. Nucleic Acids Res. *40*, 675–681.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

Stoll, D. a., Link, S., Kulling, S., Geisen, R., and Schmidt-Heydt, M. (2014). Comparative proteome analysis of *Penicillium verrucosum* grown under light of short wavelength shows an induction of stress-related proteins associated with modified mycotoxin biosynthesis. Int. J. Food Microbiol. *175*, 20–29.

Strope, P.K., Skelly, D.A., Kozmin, S.G., Mahadevan, G., Stone, E.A., Magwene, P.M., Dietrich, F.S., and McCusker, J.H. (2015). The 100-genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. Genome Res. gr.185538.114 –.

Subramanian, A.R., Hiran, S., Steinkamp, R., Meinicke, P., Corel, E., and Morgenstern, B. (2010). DIALIGN-TX and multiple protein alignment using secondary structure information at GOBICS. Nucleic Acids Res. *38*.

Szewczyk, E., Chiang, Y.-M., Oakley, C.E., Davidson, A.D., Wang, C.C.C., and Oakley, B.R. (2008). Identification and characterization of the asperthecin gene cluster of *Aspergillus nidulans*. Appl. Environ. Microbiol. *74*, 7607–7612.

Terabayashi, Y., Shimizu, M., Kitazume, T., Masuo, S., Fujii, T., and Takaya, N. (2012). Conserved and specific responses to hypoxia in Aspergillus oryzae and Aspergillus nidulans determined by comparative transcriptomics. Appl. Microbiol. Biotechnol. *93*, 305–317.

Thomas-Chollier, M., Sand, O., Turatsinze, J.V., Janky, R., Defrance, M., Vervisch, E., Brohée, S., and van Helden, J. (2008). RSAT: regulatory sequence analysis tools. Nucleic Acids Res. *36*.

Umemura, M., Koike, H., Nagano, N., Ishii, T., Kawano, J., Yamane, N., Kozone, I., Horimoto, K., Shin-ya, K., Asai, K., *et al.* (2013). MIDDAS-M: Motif-independent *de novo* detection of secondary metabolite gene clusters through the integration of genome sequencing and transcriptome data. PLoS One *8*.

Vongsangnak, W., Olsen, P., Hansen, K., Krogsgaard, S., and Nielsen, J. (2008). Improved annotation through genome-scale metabolic modeling of *Aspergillus oryzae*. BMC Genomics *9*, 245.

Wallace, I.M., O'Sullivan, O., Higgins, D.G., and Notredame, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res. *34*, 1692–1699.

Wang, F.-Q., Zhong, J., Zhao, Y., Xiao, J., Liu, J., Dai, M., Zheng, G., Zhang, L., Yu, J., Wu, J., *et al.* (2014). Genome sequencing of high-penicillin producing industrial strain of Penicillium chrysogenum. BMC Genomics 15 (Suppl. 1), S11.

Weber, T. (2014). In silico tools for the analysis of antibiotic biosynthetic pathways. Int. J. Med. Microbiol. *304*, 230–235.

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H.U., Bruccoleri, R., Lee, S.Y., Fischbach, M. a., Muller, R., Wohlleben, W., *et al.* (2015). antiSMASH 3.0 – a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. 1–7.

Wheeler, D.L., Chappey, C., Lash, A.E., Leipe, D.D., Madden, T.L., Schuler, G.D., Tatusova, T. a, and Rapp, B. a (2013). Database resources of the National Center for Biotechnology Information\. *28*, 10–14.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Prüss, M., Reuter, I., and Schacherer, F. (2000). TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res. *28*, 316–319.

Wolf, Y.I., and Koonin, E.V. (2012). A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. Genome Biol. Evol. *4*, 1286–1294.

Wood, D.E., Lin, H., Levy-Moonshine, A., Swaminathan, R., Chang, Y.-C., Anton, B.P., Osmani, L., Steffen, M., Kasif, S., and Salzberg, S.L. (2012). Thousands of missed genes found in bacterial genomes and their analysis with COMBREX. Biol. Direct *7*, 37.

Yang, Y., Zhao, H., Barrero, R. a, Zhang, B., Sun, G., Wilson, I.W., Xie, F., Walker, K.D., Parks, J.W., Bruce, R., *et al*. (2014). Genome sequencing and analysis of the paclitaxel-producing endophytic fungus Penicillium aurantiogriseum NRRL 62431. BMC Genomics *15*, 69.

Yu, J., Jurick, W.M., Cao, H., Yin, Y., Gaskins, V.L., Losada, L., Zafar, N., Kim, M., Bennett, J.W., and Nierman, W.C. (2014). Draft genome sequence of *Penicillium expansum* strain R19, which causes postharvest decay of apple fruit. Genome Announc. *2*, 2013–2014.

Bibliography

M. R. Andersen, M. P. Salazar, P. J. Schaap, P. J. Van De Vondervoort, D. Culley, J. Thykaer, J. C. Frisvad, K. F. Nielsen, R. Albang, K. Albermann, R. M. Berka, G. H. Braus, S. A. Braus-Stromeyer, L. M. Corrochano, Z. Dai, P. W. Van Dijck, G. Hofmann, L. L. Lasure, J. K. Magnuson, H. Menke, M. Meijer, S. L. Meijer, J. B. Nielsen, M. L. Nielsen, A. J. Van Ooyen, H. J. Pel, L. Poulsen, R. A. Samson, H. Stam, A. Tsang, J. M. Van Den Brink, A. Atkins, A. Aerts, H. Shapiro, J. Pangilinan, A. Salamov, Y. Lou, E. Lindquist, S. Lucas, J. Grimwood, I. V. Grigoriev, C. P. Kubicek, D. Martinez, N. N. Van Peij, J. A. Roubos, J. Nielsen, and S. E. Baker. Comparative genomics of citric-acid-producing Aspergillus niger ATCC 1015 versus enzyme-producing CBS 513.88. *Genome Research*, 21 (6):885–897, 2011. ISSN 10889051. doi: 10.1101/gr.112169.110.

T. Awakawa, X. L. Yang, T. Wakimoto, and I. Abe. Pyranonigrin E: A PKS-NRPS hybrid metabolite from aspergillus niger identified by genome mining. *ChemBioChem*, 14(16):2095–2099, 2013. ISSN 14394227. doi: 10.1002/cbic.201300430.

G. F. Bills, Q. Yue, L. Chen, Y. Li, Z. An, and J. C. Frisvad. Aspergillus mulundensis sp. nov., a new species for the fungus producing the antifungal echinocandin lipopeptides, mulundocandins. *Journal of Antibiotics*, 69(3): 141–148, 2016. ISSN 18811469. doi: 10.1038/ja.2015.105. URL `http://dx.doi.org/10.1038/ja.2015.105`.

D. Boettger and C. Hertweck. Molecular Diversity Sculpted by Fungal PKS-NRPS Hybrids. *ChemBioChem*, 14(1):28–42, 2013. ISSN 14394227. doi: 10.1002/cbic.201200624.

H. U. Böhnert, I. Fudal, W. Dioh, D. Tharreau, J.-L. Notteghem, and M.-H. Lebrun. A putative polyketide synthase/peptide synthetase from Magnaporthe grisea signals pathogen attack to resistant rice. *The Plant cell*, 16(9):2499–513, 2004. ISSN 1040-4651. doi: 10.1105/tpc.104.022715. URL `http://www.ncbi.nlm.nih.gov/pubmed/`

15319478$\delimiter"026E30F$nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC520948.

K. E. Bushley and B. G. Turgeon. Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. *BMC evolutionary biology*, 10:26, 2010. ISSN 1471-2148. doi: 10.1186/1471-2148-10-26.

C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421, 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-421. URL http://www.biomedcentral.com/1471-2105/10/421.

A. Chen, J. Frisvad, B. Sun, J. Varga, S. Kocsubé, J. Dijksterhuis, D. Kim, S.-B. Hong, J. Houbraken, and R. Samson. Aspergillus section Nidulantes (formerly Emericella): Polyphasic taxonomy, chemistry and biology. *Studies in Mycology*, pages 1–118, 2016. ISSN 01660616. doi: 10.1016/j.simyco.2016.10.001.

S. L. Clugston, S. A. Sieber, M. A. Marahiel, and C. T. Walsh. Chirality of peptide bond-forming condensation domains in nonribosomal peptide synthetases: the C5 domain of tyrocidine synthetase is a (D)C(L) catalyst. *Biochemistry*, 42(41):12095–104, oct 2003. ISSN 0006-2960. doi: 10.1021/bi035090+. URL http://www.ncbi.nlm.nih.gov/pubmed/14556641.

R. A. Coulombe and R. P. Sharma. Clearance and excretion of intratracheally and orally administered aflatoxin B1 in the rat. *Food and Chemical Toxicology*, 23(9):827–830, 1985. ISSN 02786915. doi: 10.1016/0278-6915(85)90283-2.

R. P. de Vries, J. Visser, P. Ronald, de Vries, R., and P. Aspergillus Enzymes Involved in Degradation of Plant Cell Wall Polysaccharides. *Microbiology and Molecular Biology Reviews*, 65(4):497–522, 2001. ISSN 1092-2172. doi: 10.1128/MMBR.65.4.497.

P. M. Dewick. *Medicinal Natural Products*. John Wiley & Sons, Ltd, Chichester, UK, feb 2009. ISBN 9780470742761. doi: 10.1002/9780470742761. URL http://doi.wiley.com/10.1002/9780470742761.

B. Diez, V. Ii, J. F. Martin, and J. L. Barredosll. The Cluster of Penicillin Biosynthetic Genes. *Biochemistry*, 265(27):16358–16365, 1990.

B. Dujon, D. Sherman, G. Fischer, P. Durrens, S. Casaregela, I. Lafentaine, J. De Montigny, C. Marck, C. Neuvéglise, E. Talla, N. Goffard, L. Frangeul,

M. Algie, V. Anthouard, A. Babour, V. Barbe, S. Barnay, S. Blanchin, J. M. Beckerich, E. Beyne, C. Bleykasten, A. Boisramé, J. Boyer, L. Cattolico, F. Confanioleri, A. De Daruvar, L. Despons, E. Fabre, C. Fairhead, H. Ferry-Dumazet, A. Groppi, F. Hantraye, C. Hennequin, N. Jauniaux, P. Joyot, R. Kachouri, A. Kerrest, R. Koszul, M. Lemaire, I. Lesur, L. Ma, H. Muller, J. M. Nicaud, M. Nikolski, S. Oztas, O. Ozier-Kalogeropoulos, S. Pellenz, S. Potter, G. F. Richard, M. L. Straub, A. Suleau, D. Swennen, F. Tekai, M. Wésolowski-Louvel, E. Westhof, B. Wirth, M. Zeniou-Meyer, I. Zivanovic, M. Bolotin-Fukuhara, A. Thierry, C. Bouchter, B. Caudron, C. Scarpelli, C. Gaillardin, J. Weissebach, P. Wincker, and J. L. Souciet. Genome evolution in yeasts. *Nature*, 430(6995):35–44, 2004. ISSN 00280836. doi: 10.1038/nature02579.

M. Eisendle, H. Oberegger, I. Zadra, and H. Haas. The siderophore system is essential for viability of Aspergillus nidulans: Functional analysis of two genes encoding L-ornithine N5-monooxygenase (sidA) and a non-ribosomal peptide synthetase (sidC). *Molecular Microbiology*, 49(2):359–375, 2003. ISSN 0950382X. doi: 10.1046/j.1365-2958.2003.03586.x.

B. S. E. Evans. *Nonribosomal Peptide and Polyketide Biosynthesis*. 2016. ISBN 9781493933730. doi: 10.1007/978-1-4939-3375-4.

R. Finking and M. a. Marahiel. Biosynthesis of nonribosomal peptides. *Annual review of microbiology*, 58:453–88, jan 2004. ISSN 0066-4227. doi: 10.1146/annurev.micro.58.030603.123615. URL `http://www.ncbi.nlm.nih.gov/pubmed/15487945`.

M. A. Fischbach, C. T. Walsh, and J. Clardy. The evolution of gene collectives: How natural selection drives chemical innovation. *Pnas*, 105(12): 4601–4608, 2008. ISSN 0027-8424. doi: 10.1073/pnas.0709132105.

J. C. Frisvad and T. O. Larsen. Chemodiversity in the genus Aspergillus. *Applied Microbiology and Biotechnology*, 99(19):7859–7877, 2015. ISSN 14320614. doi: 10.1007/s00253-015-6839-z.

J. E. Galagan, S. E. Calvo, C. Cuomo, L.-J. Ma, J. R. Wortman, S. Batzoglou, S.-I. Lee, M. Baştürkmen, C. C. Spevak, J. Clutterbuck, V. Kapitonov, J. Jurka, C. Scazzocchio, M. Farman, J. Butler, S. Purcell, S. Harris, G. H. Braus, O. Draht, S. Busch, C. D'Enfert, C. Bouchier, G. H. Goldman, D. Bell-Pedersen, S. Griffiths-Jones, J. H. Doonan, J. Yu, K. Vienken, A. Pain, M. Freitag, E. U. Selker, D. B. Archer, M. Á. Peñalva, B. R. Oakley, M. Momany, T. Tanaka, T. Kumagai, K. Asai, M. Machida, W. C. Nierman, D. W. Denning, M. Caddick, M. Hynes,

M. Paoletti, R. Fischer, B. Miller, P. Dyer, M. S. Sachs, S. A. Osmani, and B. W. Birren. Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. *Nature*, 438(7071): 1105–1115, 2005. ISSN 0028-0836. doi: 10.1038/nature04341. URL http://www.nature.com/doifinder/10.1038/nature04341.

A. Gallo, M. Ferrara, and G. Perrone. Phylogenetic study of polyketide synthases and nonribosomal peptide synthetases involved in the biosynthesis of mycotoxins. *Toxins*, 5(4):717–742, 2013. ISSN 20726651. doi: 10.3390/toxins5040717.

K. Hagimori, T. Fukuda, Y. Hasegawa, S. Omura, and H. Tomoda. Fungal malformins inhibit bleomycin-induced G2 checkpoint in Jurkat cells. *Biological & pharmaceutical bulletin*, 30(8):1379–83, 2007a. ISSN 0918-6158. doi: 10.1248/bpb.30.1379.

K. Hagimori, T. Fukuda, Y. Hasegawa, S. Omura, and H. Tomoda. Fungal malformins inhibit bleomycin-induced G2 checkpoint in Jurkat cells. *Biological & pharmaceutical bulletin*, 30(8):1379–83, 2007b. ISSN 0918-6158. doi: 10.1248/bpb.30.1379.

L. H. Hartwell and M. B. Kastan. Cell cycle control and cancer. *Science (New York, N.Y.)*, 266(5192):1821–8, dec 1994. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/7997877.

G. Holt and K. D. MacDonald. Isolation of strains with increased penicillin yield after hybridization in Aspergillus nidulans. *Nature*, 219(5154):636–637, 1968. ISSN 00280836. doi: 10.1038/219636a0.

V. Hubka, A. Nováková, S. W. Peterson, J. C. Frisvad, F. Sklenář, T. Matsuzawa, A. Kubátová, and M. Kolařík. A reappraisal of Aspergillus section Nidulantes with descriptions of two new sterigmatocystin-producing species. *Plant Systematics and Evolution*, 302(9):1267–1299, 2016. ISSN 16156110. doi: 10.1007/s00606-016-1331-5.

N. Khaldi and K. H. Wolfe. Evolutionary Origins of the Fumonisin Secondary Metabolite Gene Cluster in Fusarium verticillioides and Aspergillus niger. *International Journal of Evolutionary Biology*, 2011:1–7, 2011. ISSN 2090-052X. doi: 10.4061/2011/423821. URL http://www.hindawi.com/journals/ijeb/2011/423821/.

N. Khaldi, J. Collemare, M.-H. Lebrun, and K. H. Wolfe. Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi.

Genome biology, 9(1):R18, 2008. ISSN 1465-6906. doi: 10.1186/gb-2008-9-1-r18.

M. A. Klich. Aspergillus flavus: the major producer of aflatoxin. *Molecular Plant Pathology*, 8(6):713–722, nov 2007. ISSN 1464-6722. doi: 10.1111/j.1364-3703.2007.00436.x. URL http://www.ncbi.nlm.nih.gov/pubmed/20507532http://doi.wiley.com/10.1111/j.1364-3703.2007.00436.x.

A. Klitgaard, J. B. Nielsen, R. J. N. Frandsen, M. R. Andersen, and K. F. Nielsen. Combining Stable Isotope Labeling and Molecular Networking for Biosynthetic Pathway Characterization. *Analytical Chemistry*, 87(13): 6520–6526, jul 2015. ISSN 0003-2700. doi: 10.1021/acs.analchem.5b01934. URL http://pubs.acs.org/doi/abs/10.1021/acs.analchem.5b01934.

G. Koczyk, A. Dawidziuk, and D. Popiel. The distant siblings - A phylogenomic roadmap illuminates the origins of extant diversity in fungal aromatic polyketide biosynthesis. *Genome Biology and Evolution*, 7(11): 3132–3154, 2015. ISSN 17596653. doi: 10.1093/gbe/evv204.

S. Kroken, N. L. Glass, J. W. Taylor, O. C. Yoder, and B. G. Turgeon. Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15670–5, 2003. ISSN 0027-8424. doi: 10.1073/pnas.2532165100. URL http://www.pnas.org/content/100/26/15670.full.

Y. Lamboni, K. F. Nielsen, A. R. Linnemann, Y. K. Gezgin, K. Hell, M. J. R. Nout, E. J. Smid, M. Tamo, M. A. J. S. Van Boekel, J. B. Hoof, and J. C. Frisvad. Diversity in secondary metabolites including mycotoxins from strains of aspergillus section nigri isolated from raw cashew nuts from benin, west africa. *PLoS ONE*, 11(10):1–14, 2016. ISSN 19326203. doi: 10.1371/journal.pone.0164310.

D. P. Lawrence, S. Kroken, B. M. Pryor, and A. E. Arnold. Interkingdom gene transfer of a hybrid NPS/PKS from bacteria to filamentous Ascomycota. *PloS one*, 6(11):e28231, jan 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0028231. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028231.

A. P. MacCabe, H. V. Liemptll, H. Palissall, S. E. Unkless, and E. Pfeiferll. &(L-cu-Aminoadipyl)-L-cysteinyl-D-valine Synthetase from Aspergillus nidulans. *The Journal of Biological Chemistry*, 266(19):12646–12654, 1991.

M. Machida, K. Asai, M. Sano, T. Tanaka, T. Kumagai, G. Terai, K. I. Kusumoto, T. Arima, O. Akita, Y. Kashiwagi, K. Abe, K. Gomi, H. Horiuchi, K. Kitamoto, T. Kobayashi, M. Takeuchi, D. W. Denning, J. E. Galagan, W. C. Nierman, J. Yu, D. B. Archer, J. W. Bennett, D. Bhatnagar, T. E. Cleveland, N. D. Fedorova, O. Gotoh, H. Horikawa, A. Hosoyama, M. Ichinomiya, R. Igarashi, K. Iwashita, P. R. Juvvadi, M. Kato, Y. Kato, T. Kin, A. Kokubun, H. Maeda, N. Maeyama, J. I. Maruyama, H. Nagasaki, T. Nakajima, K. Oda, K. Okada, I. Paulsen, K. Sakamoto, T. Sawano, M. Takahashi, K. Takase, Y. Terabayashi, J. R. Wortman, O. Yamada, Y. Yamagata, H. Anazawa, Y. Hata, Y. Koide, T. Komori, Y. Koyama, T. Minetoki, S. Suharnan, A. Tanaka, K. Isono, S. Kuhara, N. Ogasawara, and H. Kikuchi. Genome sequencing and analysis of Aspergillus oryzae. *Nature*, 438(7071):1157–1161, 2005. ISSN 14764687. doi: 10.1038/nature04300.

M. A. Marahiel. A structural model for multimodular NRPS assembly lines. *Nat. Prod. Rep.*, 33(2):136–140, 2016. ISSN 0265-0568. doi: 10.1039/C5NP00082C. URL `http://xlink.rsc.org/?DOI=C5NP00082C`.

Y. Miyake, C. Ito, M. Itoigawa, and T. Osawa. Isolation of the Antioxidant Pyranonigrin-A from Rice Mold Starters Used in the Manufacturing Process of Fermented Foods. *Bioscience, Biotechnology, and Biochemistry*, 71 (10):2515–2521, oct 2007. ISSN 0916-8451. doi: 10.1271/bbb.70310. URL `http://www.tandfonline.com/doi/full/10.1271/bbb.70310`.

D. Moore, G. Robson, and T. Trinci. *21st Century Guidebook to Fungi*. Cambridge University Press, Cambridge, 2011. ISBN 9780511977022. doi: 10.1017/CBO9780511977022. URL `http://ebooks.cambridge.org/ref/id/CBO9780511977022`.

H. D. Mootz, D. Schwarzer, and M. A. Marahiel. Ways of assembling complex natural products on modular nonribosomal peptide synthetases. *ChemBioChem*, 3(6):490–504, 2002. ISSN 14394227. doi: 10.1002/1439-7633(20020603)3:6⟨490::AID-CBIC490⟩3.0.CO;2-N.

W. C. Nierman, A. Pain, M. J. Anderson, J. R. Wortman, H. S. Kim, J. Arroyo, M. Berriman, K. Abe, D. B. Archer, C. Bermejo, J. Bennett, P. Bowyer, D. Chen, M. Collins, R. Coulsen, R. Davies, P. S. Dyer, M. Farman, N. Fedorova, N. Fedorova, T. V. Feldblyum, R. Fischer, N. Fosker, A. Fraser, J. L. García, M. J. García, A. Goble, G. H. Goldman, K. Gomi, S. Griffith-Jones, R. Gwilliam, B. Haas, H. Haas, D. Harris, H. Horiuchi, J. Huang, S. Humphray, J. Jiménez, N. Keller, H. Khouri, K. Kitamoto,

T. Kobayashi, S. Konzack, R. Kulkarni, T. Kumagai, A. Lafton, J. P. Latgé, W. Li, A. Lord, C. Lu, W. H. Majoros, G. S. May, B. L. Miller, Y. Mohamoud, M. Molina, M. Monod, I. Mouyna, S. Mulligan, L. Murphy, S. O'Neil, I. Paulsen, M. A. Peñalva, M. Pertea, C. Price, B. L. Pritchard, M. A. Quail, E. Rabbinowitsch, N. Rawlins, M. A. Rajandream, U. Reichard, H. Renauld, G. D. Robson, S. Rodriguez De Cordoba, J. M. Rodríguez-Peña, C. M. Ronning, S. Rutter, S. L. Salzberg, M. Sanchez, J. C. Sánchez-Ferrero, D. Saunders, K. Seeger, R. Squares, S. Squares, M. Takeuchi, F. Tekaia, G. Turner, C. R. Vazquez De Aldana, J. Weidman, O. White, J. Woodward, J. H. Yu, C. Fraser, J. E. Galagan, K. Asai, M. Machida, N. Hall, B. Barrell, and D. W. Denning. Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus. *Nature*, 438(7071):1151–1156, 2005. ISSN 14764687. doi: 10.1038/nature04332.

H. J. Pel, J. H. de Winde, D. B. Archer, P. S. Dyer, G. Hofmann, P. J. Schaap, G. Turner, R. P. de Vries, R. Albang, K. Albermann, M. R. Andersen, J. D. Bendtsen, J. a. E. Benen, M. van den Berg, S. Breestraat, M. X. Caddick, R. Contreras, M. Cornell, P. M. Coutinho, E. G. J. Danchin, A. J. M. Debets, P. Dekker, P. W. M. van Dijck, A. van Dijk, L. Dijkhuizen, A. J. M. Driessen, C. D'Enfert, S. Geysens, C. Goosen, G. S. P. Groot, P. W. J. de Groot, T. Guillemette, B. Henrissat, M. Herweijer, J. P. T. W. van den Hombergh, C. a. M. J. J. van den Hondel, R. T. J. M. van der Heijden, R. M. van der Kaaij, F. M. Klis, H. J. Kools, C. P. Kubicek, P. a. van Kuyk, J. Lauber, X. Lu, M. J. E. C. van der Maarel, R. Meulenberg, H. Menke, M. a. Mortimer, J. Nielsen, S. G. Oliver, M. Olsthoorn, K. Pal, N. N. M. E. van Peij, A. F. J. Ram, U. Rinas, J. a. Roubos, C. M. J. Sagt, M. Schmoll, J. Sun, D. Ussery, J. Varga, W. Vervecken, P. J. J. van de Vondervoort, H. Wedler, H. a. B. Wösten, A.-P. Zeng, A. J. J. van Ooyen, J. Visser, and H. Stam. Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88. *Nature biotechnology*, 25(2):221–31, feb 2007. ISSN 1087-0156. doi: 10.1038/nbt1282. URL `http://www.ncbi.nlm.nih.gov/pubmed/17259976`.

L. M. Petersen, T. T. Bladt, C. Dürr, M. Seiffert, J. C. Frisvad, C. H. Gotfredsen, and T. O. Larsen. Isolation, structural analyses and biological activity assays against chronic lymphocytic leukemia of two novel cytochalasins - Sclerotionigrin A and B. *Molecules*, 19(7):9786–9797, 2014a. ISSN 14203049. doi: 10.3390/molecules19079786.

L. M. Petersen, C. Hoeck, J. C. Frisvad, C. H. Gotfredsen, and T. O. Larsen. Dereplication guided discovery of secondary metabolites of mixed biosyn-

thetic origin from Aspergillus aculeatus. *Molecules*, 19(8):10898–10921, 2014b. ISSN 14203049. doi: 10.3390/molecules190810898.

P. Pons and M. Latapy. Computing communities in large networks using random walks. *Physics and Society*, page arXiv:physics/0512106, 2005. ISSN 07403194. doi: 10.1007/11569596. URL `http://arxiv.org/abs/physics/0512106`.

I. F. Purchase and J. J. Van der Watt. Carcinogenicity of sterigmatocystin to rat skin. *Toxicology and Applied Pharmacology*, 26(2):274–281, 1973. ISSN 10960333. doi: 10.1016/0041-008X(73)90262-7.

K. Qiao, Y. H. Chooi, and Y. Tang. Identification and engineering of the cytochalasin gene cluster from Aspergillus clavatus NRRL 1. *Metabolic Engineering*, 13(6):723–732, 2011. ISSN 10967176. doi: 10.1016/j.ymben.2011.09.008.

H. Rafiei, P. Dehghan, K. Pakshir, M. C. Pour, and M. Akbari. The concentration of aflatoxin M1 in the mothers' milk in Khorrambid City, Fars, Iran. *Advanced biomedical research*, 3:152, 2014. ISSN 2277-9175. doi: 10.4103/2277-9175.137859. URL `http://www.ncbi.nlm.nih.gov/pubmed/25221755http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4162072`.

C. Rausch, T. Weber, O. Kohlbacher, W. Wohlleben, and D. H. Huson. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Research*, 33(18):5799–5808, 2005. ISSN 03051048. doi: 10.1093/nar/gki885.

C. Rausch, I. Hoof, T. Weber, W. Wohlleben, and D. H. Huson. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC evolutionary biology*, 7:78, 2007. ISSN 14712148. doi: 10.1186/1471-2148-7-78.

A. Rubinstein, Y. Lurie, L. Groskop, and M. Weintrob. Cholesterol-Lowering Eeffects of a 10 mg Daily Dose of Lovastatin in Patients with Initial Total Cholesterol Levels 200 to 240 mg/dl (5.18 to 6.21 mmol/liter). *American Journal of Cardiology*, 68(11):1123–1126, 1991.

W. G. Sorenson, J. D. Tucker, and J. P. Simpson. Mutagenicity of the tetramic mycotoxin cyclopiazonic acid. *Applied and Environmental Microbiology*, 47(6):1355–1357, 1984. ISSN 00992240.

T. Stachelhaus, H. D. Mootz, and M. A. Marahiel. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chemistry and Biology*, 6(8):493–505, 1999. ISSN 10745521. doi: 10.1016/S1074-5521(99)80082-9.

S. Suda and R. W. Curtis. Antibiotic properties of malformin. *Applied microbiology*, 14(3):475–476, 1966. ISSN 00036919.

V. Valiante, D. J. Mattern, A. Schüffler, F. Horn, G. Walther, K. Scherlach, L. Petzke, J. Dickhaut, R. Guthke, C. Hertweck, M. Nett, E. Thines, and A. A. Brakhage. Discovery of an Extended Austinoid Biosynthetic Pathway in Aspergillus calidoustus. *ACS Chemical Biology*, 12(5):1227–1234, 2017. ISSN 15548937. doi: 10.1021/acschembio.7b00003.

H. van Liempt, H. von Döhren, and H. Kleinkauf. delta-(L-alpha-aminoadipyl)-L-cysteinyl-D-valine synthetase from Aspergillus nidulans. The first enzyme in penicillin biosynthesis is a multifunctional peptide synthetase. *The Journal of biological chemistry*, 264(7):3680–4, mar 1989. ISSN 0021-9258. URL `http://www.ncbi.nlm.nih.gov/pubmed/2645274`.

J. Wang, Z. Jiang, W. Lam, E. A. Gullen, Z. Yu, Y. Wei, L. Wang, C. Zeiss, A. Beck, E. C. Cheng, C. Wu, Y. C. Cheng, and Y. Zhang. Study of malformin C, a fungal source cyclic pentapeptide, as an anti-cancer drug. *PLoS ONE*, 10(11):1–19, 2015a. ISSN 19326203. doi: 10.1371/journal.pone.0140069.

J. Wang, Z. Jiang, W. Lam, E. A. Gullen, Z. Yu, Y. Wei, L. Wang, C. Zeiss, A. Beck, E. C. Cheng, C. Wu, Y. C. Cheng, and Y. Zhang. Study of malformin C, a fungal source cyclic pentapeptide, as an anti-cancer drug. *PLoS ONE*, 10(11):1–19, 2015b. ISSN 19326203. doi: 10.1371/journal.pone.0140069.

K. J. Weissman. The structural biology of biosynthetic megaenzymes. *Nature Chemical Biology*, 11(9):660–670, 2015. ISSN 1552-4450. doi: 10.1038/nchembio.1883. URL `http://dx.doi.org/10.1038/nchembio.1883`.

A. J. Wright. The Penicillins. *Mayo Clinic Proceedings*, 74(3): 290–307, mar 1999. ISSN 00256196. doi: 10.4065/74.3.290. URL `http://www.ncbi.nlm.nih.gov/pubmed/10090000http://linkinghub.elsevier.com/retrieve/pii/S0025619611638676`.

G. Yang. A Polyketide Synthase Is Required for Fungal Virulence and Production of the Polyketide T-Toxin. *the Plant Cell Online*, 8(11):

2139–2150, 1996. ISSN 10404651. doi: 10.1105/tpc.8.11.2139. URL http://www.plantcell.org/cgi/doi/10.1105/tpc.8.11.2139.

C.-S. Yun, T. Motoyama, and H. Osada. Biosynthesis of the mycotoxin tenuazonic acid by a fungal NRPS-PKS hybrid enzyme. *Nature communications*, 6:8758, 2015. ISSN 2041-1723. doi: 10.1038/ncomms9758. URL http://www.nature.com/ncomms/2015/151027/ncomms9758/full/ncomms9758.html.

X. Zhou, X.-X. Shen, C. T. Hittinger, and A. Rokas. Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. *Molecular Biology and Evolution*, 2017. ISSN 0737-4038. doi: 10.1093/molbev/msx302. URL http://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msx302/4644721.

Chapter 2

The genomes of *Aspergillus* section *Nigri* reveal drivers in fungal speciation

In our effort to genome sequence the whole genus *Aspergillus* we started with section *Nigri*. Characterizing the genome dynamics on the section, clade and isolate level will support the development of pipelines which can describe a genus. *Aspergillus* section *Nigri* contains many species relevant to biotechnology and biomedicine — making it an ideal use case. In this study we developed new tools to describe these fungi using comparative genomics. Core-pan genome analysis and generating homologous groups of secondary metabolite gene clusters show the high genetic versatility of the section. Furthermore, we examined CAZymes to investigate the carbohydrate degrading potential of this species. Our results lets us characterize genetic diveristy over related species and establish a model of three stages of fungal speciation: acquisition, diversification and consolidation. In summary, we established pipeline will facilitate characterization of new *de novo* sequenced fungi.

The genomes of *Aspergillus* section *Nigri* reveal drivers in fungal speciation

Tammi C. Vesth (1), Jane L. Nybo (1), Sebastian Theobald (1), Jens C. Frisvad (1), Thomas O. Larsen (1), Kristian F. Nielsen (1), Jakob B. Hoof (1), Julian Brandl (1), Asaf Salamov (3), Robert Riley (3), Morten T. Nielsen (1), Ellen K. Lyhne (1), Martin E. Kogle (1), Kimchi Strasser (9), Erin McDonnell (9), Kerrie Barry (3), Alicia Clum (3), Cindy Chen (3), Matt Nolan (3), Laura Sandor (3), Alan Kuo (3), Anna Lipzen (3), Matthieu Hainaut (4,5), Elodie Drula (4,5), Adrian Tsang (9), Bernard Henrissat (4,5,6), Ad Wiebenga (7), Miia R. Mäkelä (7,8), Ronald P. de Vries (7), Igor V. Grigoriev (3), Uffe H. Mortensen (1), Scott E. Baker (2)*, Mikael R. Andersen (1)*.

1) Department of Biotechnology and Bioengineering, Technical University of Denmark, Kgs. Lyngby, Denmark
2) US Department of Energy Joint Bioenergy Institute, Berkeley, CA, USA
3) US Department of Energy Joint Genome Institute, Walnut Creek, CA, USA
4) Architecture et Fonction des Macromolécules Biologiques, (CNRS UMR 7257, Aix-Marseille University, 13288 Marseille, France.
5) INRA, USC 1408 AFMB, 13288 Marseille, France.
6) Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia
7) Fungal Physiology, Westerdijk Fungal Biodiversity Institute & Fungal Molecular Physiology, Utrecht University, Utrecht, The Netherlands
8) Department of Food and Environmental Sciences, Division of Microbiology and Biotechnology, University of Helsinki, Finland
9) Centre for Structural and Functional Genomics, Concordia University, Montreal, Quebec, Canada
*Corresponding authors

Abstract

*Aspergillus s*ection *Nigri* comprises filamentous fungi relevant to biomedicine, bioenergy, health, and biotechnology. In order to learn more about fungal speciation, as well as potential for applications in biotechnology and biomedicine, we sequenced 23 genomes *de novo*, forming a full genome compendium for the section (26 species), as well as six *A. niger* isolates. Comparative genomics of 38 fungal genomes allowed us to assess inter- and intra-species genomic variation. We predicted 17,903 CAZymes and 2,717 secondary metabolite gene clusters, which we condensed into 455 distinct families corresponding to compound classes, 49% of which are only found in single species. We performed metabolomics and genetic engineering to correlate genotypes to phenotypes, as demonstrated for the metabolite aurasperone, and by heterologous transfer of citrate production to *A. nidulans*. All analyses supported a role in speciation for secondary metabolism and regulators and allowed us to propose a three-step model for fungal speciation.

Keywords

Genomics; *Aspergillus*; *Nigri*; Primary Metabolism; Secondary Metabolism; CAZyme

Introduction

Species in genus *Aspergillus* are of broad interest to medical[1], applied[2,3], and basic research[4]. Members of *Aspergillus* section *Nigri* ("black Aspergilli") are prolific producers of native and heterologous proteins[5,6], organic acids (in particular citric acid[7–9]), and secondary metabolites

(including biopharmaceuticals and mycotoxins like ochratoxin A). Furthermore, the section members are generally very efficient producers of extracellular enzymes[10,11], they are the production organisms for 49 out of 260 industrial enzymes[12,13]. Among the most important of these, in addition to *A. niger*, are *A. tubingensis*, *A. aculeatus*, and *A. luchuensis* (previously *A. acidus, A. kawachii,* and *A. awamori*[14–16]).

Members of *Aspergillus* section *Nigri* are also known as destructive degraders of foods and feeds, and some isolates produce the potent mycotoxins ochratoxin A[17] and fumonisins[18–20]. In addition, some species in this section have been proposed to be pathogenic to humans and other animals[21]. It is thus of interest to further examine section *Nigri* for industrial exploitation, prevention of food spoilage, toxin production, and pathogenicity caused by these fungi.

A combined phylogenetic and phenotypic approach has revealed that section *Nigri* contains at least 27 species[22–26]. Recent results have shown that the section contains species with a high diversity and may consist of two separate clades: the biseriate species and the uniseriate species[27], which show differences in sexual states[28], sclerotium formation[29], and secondary metabolite production[30]. In the section, only six species have had their genome sequenced: *A. niger*[31,32], *A. luchuensis*[16,33], *A. carbonarius*[34], *A. aculeatus*[34], *A. tubingensis*[34], and *A. brasiliensis*[34].

This section, with its combination of fungal species with diverse impact on humanity and species-richness, is thus an interesting target for studying how fungi diversify into species. In this study, we have *de novo* sequenced the genomes of 20 species of section *Nigri,* thus completing a genome compendium of 26 described species in the section. Further, we have genome-sequenced three additional *A. niger* isolates (including two previously described as species *A. lacticoffeatus*[11] and *A. phoenicis*[35]), which in combination allows for inter- and intra-species comparison of 32 isolates. The development of methods for comparative genomics, combined with experimental analysis of the species, allows us to track genetic diversity across genomes, from the protein level, over the evolution of biosynthetic gene clusters, to the groups of genes which define clades or individual species. This analysis in conjunction with the high resolution in genome sequences allows us to propose a hypothesis on fungal species evolution and diversification.

Results and Discussion

Analyzing 23 new genomes reveals high genetic diversity of *Aspergillus* section *Nigri*

We present 23 whole genome sequences: 20 genomes of section *Nigri* species previously unsequenced and three additional *A. niger* genomes for assessment of intraspecies diversity. All genomes were sequenced, assembled, and annotated using the JGI fungal genome pipeline[36,37] (SI 1). Figure 1 shows a phylogenetic tree as well as gene richness, number of scaffolds, and functional annotation (InterPro[38,39]). The tree supports previous proposals[11,35] that *A. lacticoffeatus* and *A. phoenicis* are indeed synonyms of *A. niger*.

In comparing key statistics of the genomes, some traits are quite similar, and others surprisingly variable. Many of the investigated species have around the average number of genes (11,900), but there is considerable variation from the smallest number of predicted genes (10,066) to the highest (13,687). The smallest number of predicted genes in section *Nigri* is found in *A. saccharolyticus* which supports the previous observation[40,41] that this species is quite atypical in section *Nigri*.

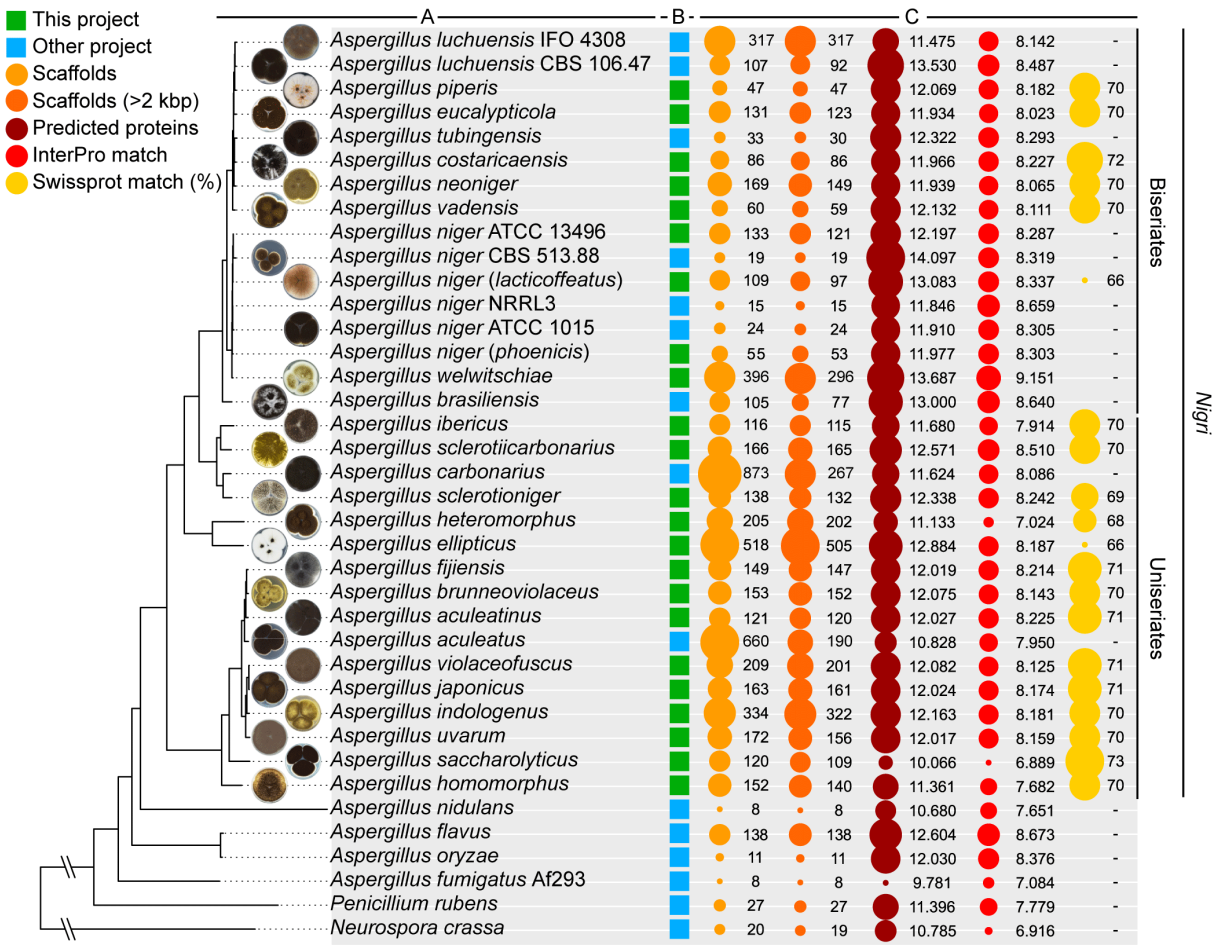


Figure 1. Dendrogram and bubble plots illustrating phylogenetic distances between 32 genomes from section *Nigri* as well as four non-*Nigri Aspergillus* species, a *Penicillium* and a *Neurospora* genome (for outgroups). Additional information is available in SI 1. A) Phylogenetic tree created using RAxML[51], MAFFT[52], and Gblocks[53], based on 2,022 conserved genes. Plate growth pictures are presented for each newly sequenced species. B) Colors indicate whether the organism is from this or another sequencing project. C) Five bubble plots of descriptive numbers for each genome. The bubble sizes have been scaled to the categories and are not comparable across categories.

We further evaluated the annotation of the 23 genome sequences we generated. The percentage of complete genes (including start and stop codon) is in the range of 94-98%, and 67% of the proteins could be assigned one or more InterPro domains. The number of scaffolds (average 166) vary from 47 in *A. piperis* to 518 in *A. ellipticus*. On average, 70% of the proteins had sequence homologs in Swissprot (91% of proteins have homologs within section *Nigri*, see next section). This means that even though six members of section *Nigri* have already been sequenced, ~30% of each of the new genomes are not found in Swissprot, which demonstrates the large genetic diversity in the section.

Constructing the pan- and core-genome of section *Nigri* shows genome flexibility and many species-unique genes

Given the genetic diversity in section *Nigri*, we were interested in examining the extent and timing of the genetic diversification events. For this analysis, we focused on three conceptual groups of genes: 1) The pan-genome: all genes present in one or more species. 2) The core-genome: genes present in all included species including paralogs. This set is expected to encode cellular functions needed

for all species. 3) Species-unique genes: found in only one species with or without paralogs. These genes are expected to be involved in environmental adaptation and speciation.

We first identified orthologs and paralogs using a BLASTP-based pipeline with reciprocal cut-offs specific to the *Aspergillus* genus (SI 2). Groups of homologous proteins are referred to as families. Figure 2A–C shows the overall genetic diversity between 38 fungal species from closely related genera, 36 species within the *Aspergillus* genus and 32 species from section *Nigri*.

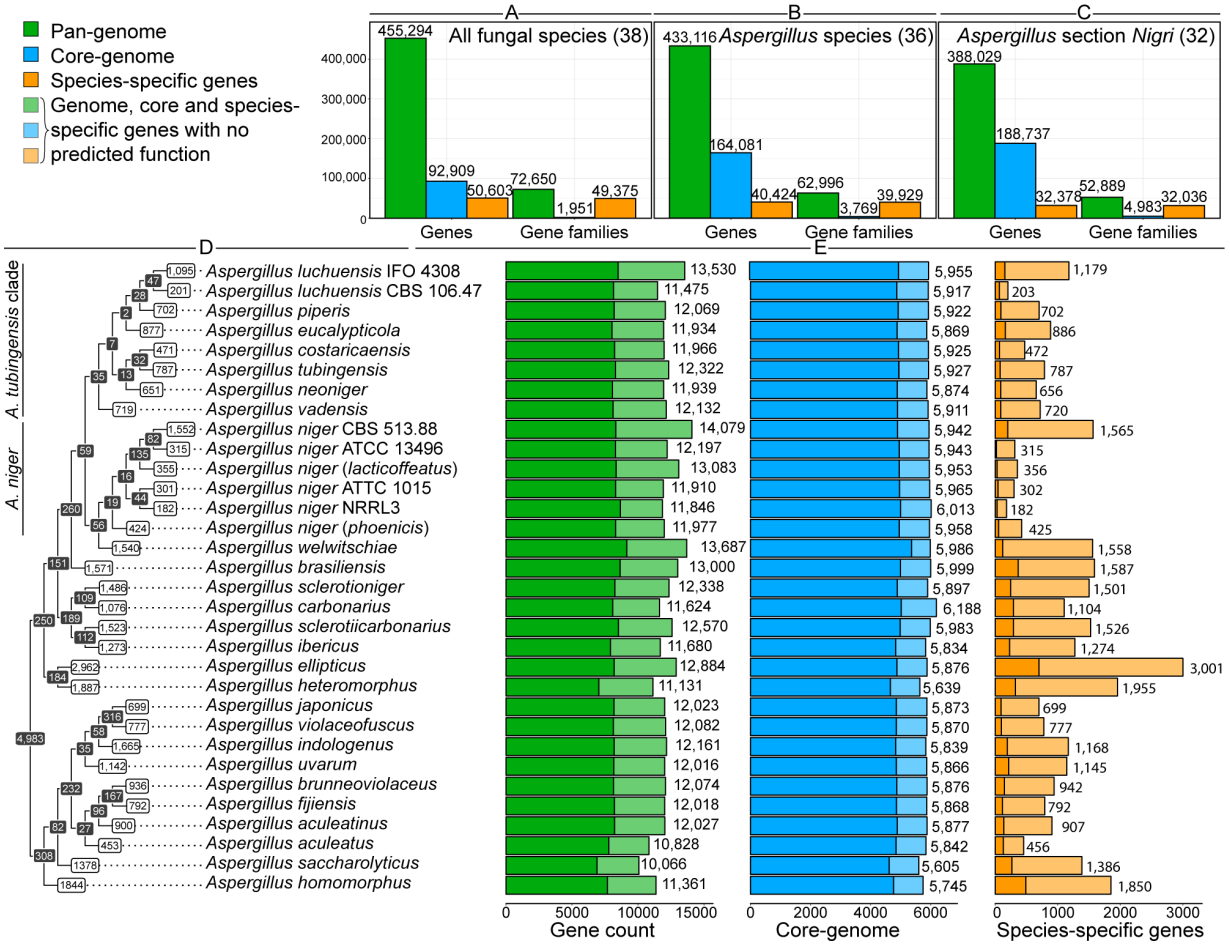


Figure 2. Genetic diversity between fungal species from closely related genera, species within the *Aspergillus* genus and section *Nigri*. (A-C) Histogram representation of the total number of genes and families distributed over the pan (green), core (blue), and unique (orange) genome for (A) the 38 fungal genomes of this study, (B) the 36 *Aspergillus* genus genomes and, (C) the 32 genomes in section *Nigri*. (D) Dendrogram of the phylogenetic relation between the 32 species in section *Nigri*. The black nodes represent homolog-families found only in the species branching from the nodes. The white boxes represent the genes unique to the specific species. (E) Stacked histograms of the gene count, the core-genome as well as unique genes for each species, numbers show total number of genes. (full colours: InterPro annotation, light colours: no annotation).

The *Aspergillus* genus pan-genome comprises 433,116 genes across the 36 *Aspergillus* genomes and from this, 62,996 gene families were constructed. 6% of those are found in all genomes (3,769 core families) while 9% are genes without orthologs in the other genomes (40,424 unique genes, 39,929 unique families) (Figure 2B). We also found evidence of potential gene transfers between species of this section, as 23% of the pan-gene families are not present in groups of species fitting the phylogenetic tree, hence indicating transfer of genes from one branch to another.

We further performed an analysis defining the number of core-gene families in section *Nigri* and in all sub-clades hereof (Figure 2D). The core-genome of section *Nigri* was found to be 32% larger than

that of the genus (4,983 families relative to 3,769, Figure 2B–C). Conversely, 9% are unique to a specific species (32,378 unique genes in 32,036 families, Figure 2C). The fraction of genes unique to a species is similar within the section and across the genus, meaning that adding a new section *Nigri* genome adds as many new genes as adding a more distantly related *Aspergillus.* This is rather interesting and shows a generally high genetic diversity of genus *Aspergillus*. Such a tendency could be the result of over-predicting genes, considering the low rate of InterPro annotation in the unique genes (Figure 2E); however, it could also be the result of horizontal gene transfer from uncharacterized species.

The section *Nigri* core genome contains signature genes of the black Aspergilli

To associate biological functions to the pan-, core-, and unique genomes, and genes exclusive to only members of the black *Aspergilli*, we employed the InterPro database[39]. Examining the core-genome of 38 fungal genomes (Figure 2A), only 4.5% of the genes lack InterPro domains (SI 3, 4.1), indicating – as would be expected – that the core-genes across closely related fungal genera include generally known and conserved functions. For the pan-genome of the 36 *Aspergillus* species versus the section *Nigri* pan genome, the percentage of unknown function is similar (32% vs 33%, SI 4.4, 6.1), as are the numbers for the core-genomes (14% vs 17%, Figure 2E, SI 4.4, 6.1). General functions like transporters, regulators, organelle specific proteins, primary metabolism, and structural domains were found as core-features across all 36 Aspergilli (SI 4.6), which supports the general validity of the method.

We expected the section *Nigri* core-genome (gene families only found in all the species of section *Nigri* and not in other Aspergilli examined) to contain *Nigri* signature genes, and found this to be the case. These families contain 580 InterPro domains conserved to a varying degree, including a high number of genes involved in the saprotropic lifestyle and secondary metabolism (SI 7).

Unique genes involved in secondary metabolism and regulation appear to be common drivers of speciation

The genetic diversity seen in section *Nigri* led us to investigate whether the unique genes for each species could provide clues to the cause of fungal speciation. While these genes by definition do not have homologs in other species, we can predict general functions using InterPro domains. Unique genes of species in section *Nigri* matched 1,334 different InterPro domains (SI 8.3). Within this, we searched for common functions in all genes unique to section *Nigri* species (excluding six *A. niger* isolates due to intraspecies variation). Surprisingly, we only identified ten domains which were found in nearly all *Nigri* species (25–26 species). Notably, nine in ten are related to functions involved in secondary metabolites or gene or protein regulation (SI 9). Finding these functions in nearly all sets of species-specific genes suggests that secondary metabolite production and regulatory proteins are drivers in fungal speciation.

Intra-species genetic variation is similar to inter-species variation

We were interested in comparing the diversity between isolates of the same species to that of the diversity of species in the same clade. We thus compared six *A. niger* isolates to the eight closely related species in the *A. tubingensis* clade (Figure 2D). The *A. niger* isolates have a high degree of genetic homogeneity as 80% of the *A. niger* pan-genome is conserved across the six isolates and only 6% is unique to any of the isolates (SI 10A). The same scale is seen in the *A. tubingensis* clade (77% shared pan-genome, 7% unique, SI 10B). Moreover, the percentage of genes with predicted functional domains within the two groups is similar to that of section *Nigri* as a whole (SI 11, 6.1, 12.1, 12.4). For the unique genes of the two groups, these are largely of unknown function (*A.*

tubingensis clade 82%, *A niger* complex 86%, SI 12.1, 12.4, 13.1, 13.2). The functions of the *A. niger* core-genome (3,798 domains) are, not surprisingly, very similar to that of section *Nigri* as a whole (SI 6.3, 12.3). In summary, the inter-species variation in the *A. tubingensis* clade is of the same scale as the intra-species variation in the *A. niger* isolates, showing that a large genetic variation does not directly translate to the currently circumscribed species.

Acid-producing species have extra citrate synthase genes which confer increased citrate production in *A. nidulans*

As species of section *Nigri* are known organic acid-producers, the genes involved in central metabolism are of interest; as the cause of citric acid overproduction in several of the section members is still not identified[42]. We thus analysed the number of paralogs in central carbon metabolism in our set of 38 fungal genomes using a curated version of an *A. niger* genome-scale metabolic model[43] as a source of pathway annotation (Figure 3).

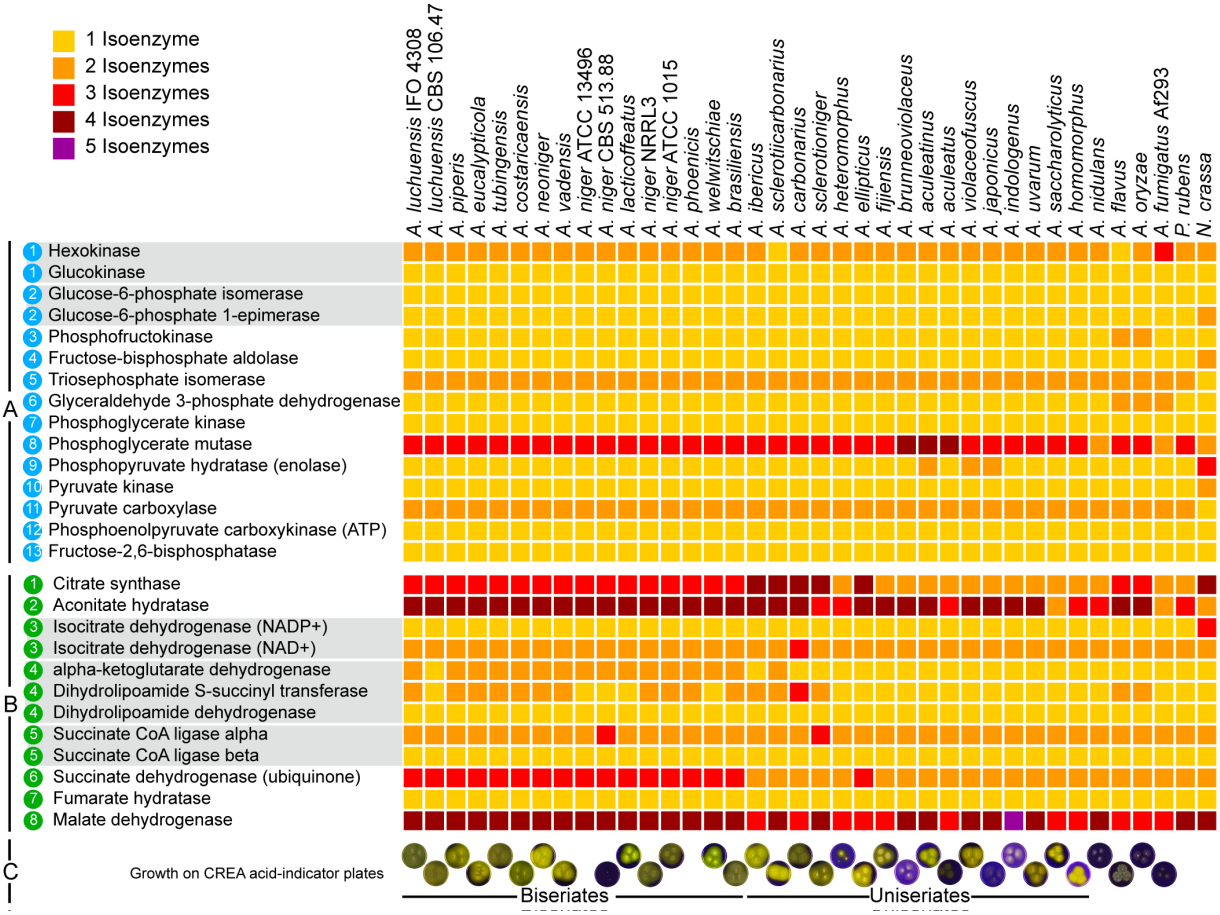


Figure 3. Gene family sizes for 31 genomes from section *Nigri* and six reference species. Proteins putatively catalyzing the same metabolic step (i.e. complex subunits or functionally related enzymes) have been grouped in grey boxes. Species are sorted according to the phylogeny of Figure 1. A) Glycolysis isoenzymes. B) TCA cycle isoenzymes. C) Growth of strains on CREA medium, which turns yellow with lowered pH. Note that *A. niger* CBS 513.88 has been mutagenized[31].

The analysis of paralogs in glycolysis shows very little variance across the 32 *Nigri* genomes but varies compared to the six other fungal species (Figure 3A). For the TCA cycle (Figure 3B), it is evident that certain metabolic steps in the pathway are conserved throughout all species, while

others vary in paralog numbers. The biseriates are particularly homogeneous. These, along with four uniseriates, are also the primary citric acid-producing species in the section (Figure 3C).

Of particular interest is the citrate production phenotype, and thus citrate synthase. All biseriates have one extra citrate synthase, the four acid-producing uniseriates have two extra. Sequence alignment revealed three distinct types (SI 14), two of which are mitochondrial and found in all species. All extra citrate synthase paralogs are of the third type, predicted to be cytosolic. We identified the extra biseriate citrate synthase (*citB*[44]) in a conserved 30 kB gene cluster including two transcription factors, a transporter and two putative fatty acid synthases. We performed heterologous expression of the *A. niger citB* gene cluster in *A. nidulans* (which only has the two mitochondrial citrate synthases) using two constitutive promoters to control the transcription factors. This expression increased citrate concentrations by 42-52% (SI 15). We hypothesize that this gene cluster may have a particular role in citrate production and additional undescribed functions involving the fatty acid synthase-like genes.

Species-specific carbon utilization is not correlated with CAZyme content, but may be determined by CAZyme regulation

Aspergilli have a particularly broad ability to degrade and convert plant biomass[34]. It is thus essential to examine the species diversity of this trait. We predicted the CAZy gene content of the genomes across section *Nigri* (17,903 CAZy domains, Figure 4, SI 16) and performed growth profiling on plant biomass-related carbon sources (SI 17). Growth on D-glucose was used to evaluate relative growth, showing variation between species.

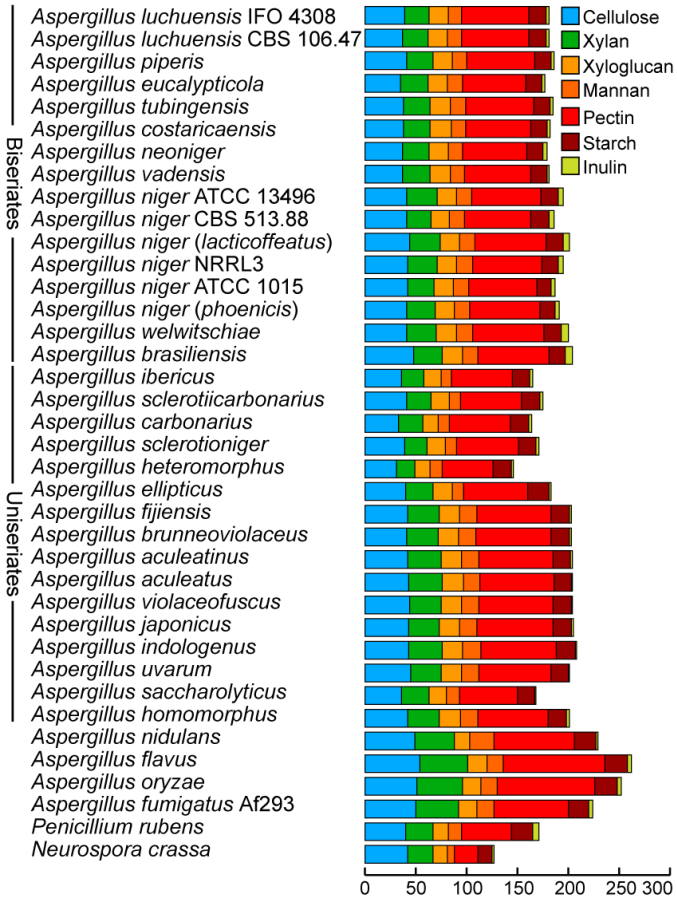


Fig. 4. Comparison of CAZy gene content divided by target polysaccharide. Details on CAZy families are available in SI 16. Growth profiles are available in SI 17. 70

All black Aspergilli grew well on pectin and have a highly conserved and extensive set of genes encoding pectin-active enzymes. Growth on other plant polysaccharides such as xylan, starch, and guar gum is more variable, despite highly conserved genes related to xyloglucan and starch degradation. The growth and genetic variability on inulin is particularly high, nine species showed reduced growth. Moreover, endoinulinase (GH32 INU) is only present in eight of the black Aspergilli while the remainder of inulin-related genes (GH32 INV and INX) are more commonly present (SI 16). However, the growth phenotypes show no correlation with the gene content (SI 17).

In a previous study[11], enzyme levels were measured in several black Aspergilli and significant differences were found. However, these differences do not reflect the genetic differences seen here (SI 16). Considering the relative uniformity of the CAZyme content (Figure 4), no correlation between genome content and growth on plant biomass-related carbon sources (SI 17) was observed for the black Aspergilli, suggesting that the differences in the capability of plant biomass degradation reflect gene expression levels in the individual fungus. This confirms a proteome study of less-related Aspergilli where the different response to plant biomass appeared to be mainly at the regulatory level[45]. The data suggest that this is the case for section *Nigri:* Species-specific phenotypes is not generally driven by CAZyme content in closely related species, but by regulation.

Section-level analysis of secondary metabolism identifies 2,717 gene clusters in 455 distinct families

Secondary metabolism (SM) is thought to be the basis of chemical defence, virulence, toxicity, mineral uptake and communication in fungi[46] and have a wide range of potential medical applications. As it was also suggested above to be involved in speciation, we examined the exometabolite diversity of 37 *Aspergillus* and *Penicillium* species based on predictions of SM gene clusters (SMGCs) as well as chemical profiles of the species of section *Nigri* on multiple substrates.

We identified 2,717 SMGCs in the 37 genomes. This is an even higher number of SMGCs per species than a previous study found in 24 *Penicillium* genomes[47]. We were further interested in quantifying the actual diversity of the SMGCs in section *Nigri* and to analyse presence patterns of SMGCs across species. We therefore defined SMGC “families” as genetically similar SMGCs across genomes (SI 2). Each SMGC family is expected to produce the same or similar compounds. This clustering resulted in the definition of 455 SMGC families across the 37 genomes (SI 19), indicating the potential production of 455 different chemical families. Most (82%) are found in less than 10 organisms, and 49% contain only one cluster (SI 20 shows examples). This is on average 8.75 unique clusters per species, despite the close phylogenetic distance of the section.

Phylogenetic examination of secondary metabolite families shows dynamics and supports involvement in speciation

To understand more about how SMGCs are exchanged between species, each of the 455 SMGC families was characterized by the type of backbone enzyme and analysed according to the phylogeny (Figure 5A–B). Only five out of all SMGCs were present in all analysed species including clusters for the NRPS-derived siderophore ferrichrome, the circular NRP fungisporin[48]/nidulanin A[49] and pigment (*YWA1*) synthesis. Two shared SMGC families were false predictions, namely two fatty acid synthases.

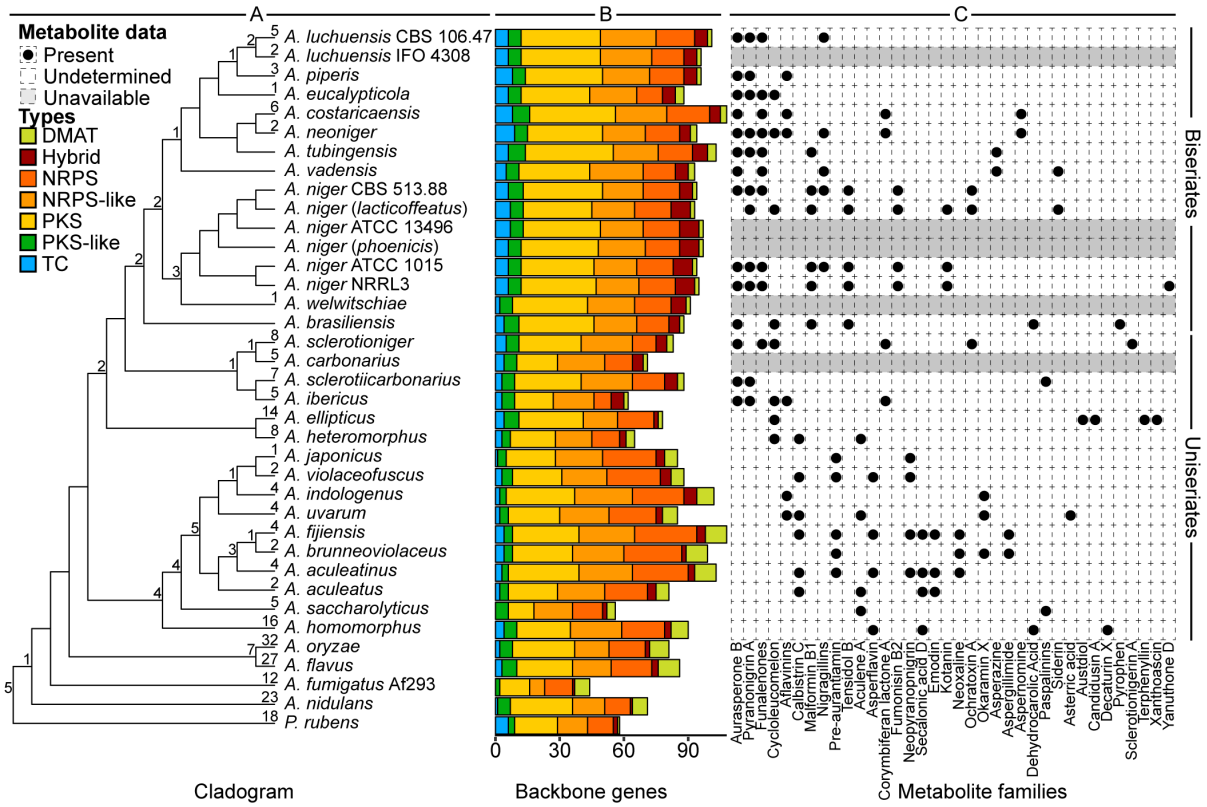


Figure 5. Shared secondary metabolism gene clusters (SMGC) and types of cluster backbones for section *Nigri*. A) Cladogram of 36 *Aspergillus* (32 *Nigri*) genomes and one *Penicillium* genome. Numbers at nodes show SMGC families shared by species in the branch. End points show families unique to an organism. B) Total SMGC family profile of the organism classified by backbone enzyme type. C) Presence/absence map of detected secondary metabolites in section *Nigri* species (SI 2). Strains with no available metabolite data are marked in light gray. DMAT: Dimethylallyltransferase (Prenyl transferases), HYBRID: Gene containing domains from NRPS and PKS backbones, NRPS: Non-ribosomal peptide synthetase, NRPS-like: Non-ribosomal peptide synthetase like containing at least two NRPS specific domains and another domain or one NRPS A domain in combination with NAD_binding_4 domain or short chain dehydrogenase., PKS: Polyketide synthase, PKS-like: Polyketide synthase like, containing at least two PKS specific domains and another domain, TC: Terpene cyclase.

Examining the dynamics of the families, only 32% and 19% of SMGCs found in two or three organisms respectively follow the whole-genome phylogeny and suggest intra-genus horizontal gene transfer (HGT) or SMGC loss to be relatively common. As an example, an SMGC is found in five distantly related species (SI 20B).

As seen in Figure 5A, the presence of unique SMGC families at every major branch point in the phylogenetic tree supports that SMGC evolution is a driver in speciation: All biseriates share a previously undescribed PKS and an NRPS-like, the *A. carbonarius* clade a terpene cyclase, and the remaining biseriates share another PKS and an NRPS-like. Furthermore, the *A. niger* complex and *A. tubingensis* clade each share unique PKS genes. Uniseriates share four unique previously undescribed SMGC families (Figure 5A). Examining the individual species, every single section Nigri species with the exception of *A. tubingensis* has a unique set of SMGCs. Conversely, all six *A. niger* isolates all have the same set of SMGCs. These patterns show the high diversity of secondary metabolism as well as support the role of these genes in speciation.

Correlation of the exometabolome with SMGC families pinpoints SMGC candidates

As a further application of the constructed SMGC families, we hypothesized that we can correlate SMGC families to classes of compounds. We performed extensive exometabolome analysis of 27 of the sequenced strains, and identified 35 compound families (Figure 5C, SI 21).

The most abundant group was naphtha-gamma-pyrones, of which aurasperone B[30] was identified in 14 of the isolates. We compared the presence patterns of SMGC families with the compound class and combined it with a knowledge-based filtering of InterPro domains leaving one hit (SI 20D). The candidate SMGC family is a nine-gene cluster found in 18 genomes – including the 14 where we detect the compound – and it contains all activities needed to synthesize aurasperone. As a support of this identification, a SMGC for a closely related compound, aurofusarin, has been experimentally verified in *Fusarium graminearium*[50]. The aurasperone cluster shares six genes (whereof one is a duplication) with the aurofusarin cluster. This supports the assignment of this family of SMGCs to the production of aurasperone B, and conceptually supports this approach for efficient linking of clusters to compounds. We see this correlation approach as highly useful for future elucidations of fungal metabolites.

A proposed model for fungal speciation

Based on our analysis above, we note that both secondary metabolites and regulatory genes would have a phenotype in the host right after acquisition, making them more likely to drive diversification and ultimately speciation. We speculate that this happens in three steps: 1) Acquisition: One or more genes provide superior fitness or a competitive advantage. 2) Diversification: A growth advantage allows the genome to diverge without strong selective pressure. The diverse *A. niger* genomes could be seen as being in this stage (Figures 1, 2, and 5). 3) Consolidation: Traits are fixed in the population, and a species with a new lifestyle can proliferate (as seen in the nodes of Figure 2D and the SMGC analysis below).

Conclusion

We have sequenced the genomes of a whole section of filamentous fungi, and a diverse set of *A. niger* isolates, and found that the species are highly diverse on some traits, in particular secondary metabolism, and homogeneous on others, such as glycolytic metabolism, and CAZymes. The presented data furthermore provides an extensive compendium of 24 new genomes which provides substantial information on fungal genetic diversity. We further combined genome analysis with metabolite profiling and genetic engineering to identify the genetic basis of several phenotypes within primary and secondary metabolism.

Of particular interest was the finding that the species-specific genes in all species share functions within gene/protein regulation and secondary metabolism. This suggests that these highly transferable traits are particularly active and efficient drivers of fungal speciation.

Acknowledgements

TCV, JLN, ST, and MRA acknowledge funding from The Villum Foundation, grant VKR023437. JB and MRA acknowledge funding from the Novo Nordisk Foundation, grant NNF13OC0004831. MRA and TCV acknowledge funding from the Danish National Research Foundation (DNRF137) for the Center for Microbial Secondary Metabolites. JCF acknowledge funding from the Novo Nordisk Foundation, grant NNF13OC0005201. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Inge Kjaerboelling is acknowledged for critical and constructive feedback on the manuscript and figures.

Supplementary Information

SI 1

Genome statistics of sequencing and annotation. Absence of information is indicated by a dash. Details for column contents are as follows: nrHits: proteins which had a significant hit in Swissport or NCBI non-redundant database. completeCDS : gene predictions containing both start and stop codon. 75percentCovByDNA: genes with >= 75% coverage of exons by RNA. totalExonsCovByRNA: genes with 100% coverage of exons by RNA. totalGenes: total number of predicted genes. repeatRegions: number of Repeat-covered regions. length_repeatRegions: total length of repeat-covered regions. seqType: sequencing setup. assemblyType: assembly pipeline. study: genome sequenced in this or other studies.

SI 2

Materials, Methods, and Protocols

SI 3

Dendrogram and InterPro coverage of the 36 *Aspergillus* genus species A. Dendrogram of the phylogenetic relation between the 36 species. The black boxes represent the homologous genes among the species branching from the nodes. The white boxes represent the genes unique to the specific species. B. Genome sizes of each species. The black coloured boxes represent the genes annotated by InterPro. C. Core-genome size of each species. The blue coloured boxes represent the genes annotated by InterPro. D. Species unique genes. The yellow coloured boxes represent the genes annotated by InterPro. The grey coloured boxes represent genes with no annotation. The numbers at right side of the boxes indicates the total number of annotated and not annotated genes. The bar scales are unique to each graph.

SI 4

Sheet 1: InterPro coverage of the 38 fungal species. The genome size and the InterPro coverage of each of the 38 fungal species alongside the core and species unique segments of the genome. The gene number representing the core and unique portion of the genomes will adjust relative to the accompanying species. Sheet 2: InterPro domains and functions included in the pan-genome of the 38 fungal species. Sheet 3: InterPro domains in the core-genome of the 38 fungal species.
Sheet 4: InterPro coverage of the 36 *Aspergillus* genus species. The genome size and the InterPro coverage of each of the 36 *Aspergillus* genus alongside the core and species unique segments of the genome. The gene number representing the core and unique portion of the genomes will adjust relative to the accompanying species. Sheet 5: InterPro domains and functions included in the pan-genome of the 36 *Aspergillus* genus species. Sheet 6: InterPro domains in the core-genome of the 36 *Aspergillus* genus species.

SI 5

Phylogenetic relation and InterPro coverage of the fungal species included in this study. A. Dendrogram of the phylogenetic relation between the 38 species. The black boxes represent the homologous genes among the species branching from the nodes. The white boxes represent the genes unique to the specific species. B. Genome sizes of each species. The black coloured boxes represent the genes annotated by InterPro. C. Core-genome size of each species. The blue coloured boxes represent the genes annotated by InterPro. D Species unique genes. The yellow coloured boxes represent the genes annotated by InterPro. The grey coloured boxes represent genes with no

annotation. The numbers at right side of the boxes indicates the total number of annotated and not annotated genes. The bar scales are unique to each graph.

SI 6

Sheet 1: InterPro coverage of section *Nigri* species. The genome size and the InterPro coverage of each of section *Nigri* species alongside the core and species unique segments of the genome. The gene number representing the core and unique portion of the genomes will adjust relative to the accompanying species. Sheet 2: InterPro domains and functions included in the pan-genome of section *Nigri* species. Sheet 3: InterPro domains in the core-genome of the section *Nigri* species.

SI 7

The InterPro domains for the core-genome specific to section *Nigri*. The InterPro domains not found in the core-genome of the non-*Nigri Aspergillus* species.

SI 8

InterPro domains in the unique gene set. Sheet 1: InterPro domains in the unique gene set of the 38 fungal species in this study. Sheet 2: InterPro domains in the unique gene set of the *Aspergillus* genus. Sheet 3: InterPro domains in the unique gene set of section *Nigri*.

SI 9

InterPro domains potentially involved in general speciation within section *Nigri* without *A.* niger isolates. InterPro domains in the unique gene set of section *Nigri* without the *A. niger* isolates where the InterPro domain is found in 25 or 26 species out of 26 species.

SI 10

A. Pan, core, unique genes and families of the *Nigri* clade species
The total number of the proteins and families of all genes in the Tubingensis clade (pan-genome, green), genes shared by all species (core-genome, blue), and genes unique to the individual species (species unique genes, orange). The families were predicted using BLASTp alignments with cutoffs specific to the *Aspergillus* genus and single linkage clustering designed for this project. *Nigri* clade species: *A. lacticoffeatus, A. niger ATCC 1015, A. niger ATCC 13496, A. niger CBS 513.88, A. niger NRRL3*, and *A. phoenicis*.
B. Pan, core, unique genes and families of the Tubingensis clade species
The total number of the proteins and families of all genes in the Tubingensis clade (pan-genome, green), genes shared by all species (core-genome, blue), and genes unique to the individual species (species unique genes, orange). The families were predicted using BLASTp alignments with cutoffs specific to the *Aspergillus* genus and single linkage clustering designed for this project. Tubingensis clade species: *A. costaricaensis, A. eucalypticola, A. luchuensis CBS 106.47, A. luchuensis IFO 4308, A. neoniger, A. piperis, A. tubingensis* and *A. vadensis.*

SI 11

Phylogenetic relation and InterPro coverage of the *A. tubingensis* clade species. A. Dendrogram of the phylogenetic relation between the 8 species. The black boxes represent the homologous genes among the species branching from the nodes. The white boxes represent the genes unique to the specific species. B. Genome sizes of each species. The black coloured boxes represent the genes annotated by InterPro. C. Core-genome size of each species. The blue coloured boxes represent the genes annotated by InterPro. D Species unique genes. The yellow coloured boxes represent the genes annotated by InterPro. The grey coloured boxes represent genes with no annotation. The

numbers at right side of the boxes indicates the total number of annotated and not annotated genes. The bar scales are unique to each graph.

Dendrogram and InterPro coverage of the *A. niger* clade species

Phylogenetic relation and InterPro coverage of the *A. niger* clade species. A. Dendrogram of the phylogenetic relation between the 6 species. The black boxes represent the homologous genes among the species branching from the nodes. The white boxes represent the genes unique to the specific species. B. Genome sizes of each species. The black coloured boxes represent the genes annotated by InterPro. C. Core-genome size of each species. The blue coloured boxes represent the genes annotated by InterPro. D Species unique genes. The yellow coloured boxes represent the genes annotated by InterPro. The grey coloured boxes represent genes with no annotation. The numbers at right side of the boxes indicates the total number of annotated and not annotated genes. The bar scales are unique to each graph.

SI 12

Sheet 1: InterPro coverage of the Niger clade species. The genome size and the InterPro coverage of each of the Niger clade species alongside the core and species unique segments of the genome. The gene number representing the core and unique portion of the genomes will adjust relative to the accompanying species. Sheet 2: InterPro domains and functions included in the pan-genome of the Niger clade species. Sheet 3: InterPro domains in the core-genome of the Niger clade species. Sheet 4: InterPro coverage of the Tubingensis clade species. The genome size and the InterPro coverage of each of the Tubingensis clade alongside the core and species unique segments of the genome. The gene number representing the core and unique portion of the genomes will adjust relative to the accompanying species. Sheet 5: InterPro domains and functions included in the pan-genome of the Tubingensis clade species. Sheet 6: InterPro domains in the core-genome of the Tubingensis clade species.

SI 13

InterPro domains in the unique gene set. Sheet 1: InterPro domains and functions in the Niger clade unique gene set. Sheet 2: InterPro domains and functions in the Tubingensis clade unique gene set.

SI 14

Sequence comparison of citrate synthase genes. The sequence of the putative citrate synthases identified in the copy number analysis has been compared using neighbor-joining as implemented in MUSCLE. Potential subcellular locations have been predicted using the TargetP method.

SI 15

Overview of areas under curve for citrate production using HPLC quantification. Citrate was measured in duplicate experiments for three strains cultivated with and without manganese added to the growth medium.

SI 16

Table providing overviews of number of Glycoside Hydrolases (GH), Glycosyl Transferases (GT), Polysaccharide Lyases (PL), Carbohydrate Esterases (CE), Carbohydrate Binding Modules (CBM) and Auxiliary Activities (AA) in 37 genomes included in the study.

SI 17

Comparative growth profiling of the aspergilli from section *Nigri* and reference fungi.

SI 18

Schematic representation of secondary metabolic gene cluster family identification. Protein blast comparisons between all gene cluster members are aggregated into one cluster similarity score. From this, a network is created with gene clusters as nodes (dots) and similarity score as edges (arrows). Subsequently, random walk clustering is used to find communities of nodes inside the network. Green arrows indicate a high probability that nodes will be assigned to one community. Red arrows indicate community borders. Resulting families are containing communities of related SMGC.

SI 19

Barplot describing SMGC family frequencies. The bars illustrate the presence of a SMGC family in a certain number of organisms. Numbers above bars show the total counts.

SI 20

Overview of predicted SMGC families (B-D) and their location in the phylogeny. A: Cladogram of species used for secondary metabolic gene cluster analysis in this study. Letter code indicates predicted clusters for fumonisin, aurasperone B and an example gene cluster family predicted exclusively in biseriates. B: Example gene cluster found in five distantly related species**.** C: Predicted gene cluster family containing a PKS in *Aspergillus niger* CBS 513.8 similar to *FUM 1* from *Fusarium oxysporum*. D: Predicted gene cluster family for aurasperone B including the aurofusarin gene cluster from *Fusarium graminearum*[50]

SI 21

Overview of SM families detected in 27 different *Aspergillus* isolates

References

1. Nierman, W. C. *et al.* Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature* **438,** 1151–6 (2005).
2. Pel, H. J. *et al.* Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. *Nat. Biotechnol.* **25,** 221–31 (2007).
3. Machida, M. *et al.* Genome sequencing and analysis of *Aspergillus oryzae*. *Nature* **438,** 1157–61 (2005).
4. Galagan, J. E. *et al.* Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* **438,** 1105–1115 (2005).
5. Papagianni, M. Advances in citric acid fermentation by *Aspergillus niger*: biochemical aspects, membrane transport and modeling. *Biotechnol Adv* **25,** 244–263 (2007).
6. Punt, P. J. *et al.* Filamentous fungi as cell factories for heterologous protein production. *Trends in Biotechnology* **20,** 200–206 (2002).
7. Currie, J. The citric acid fermentation of *Aspergillus niger*. *J. Biol. Chem.* **31,** 15–37 (1917).
8. Pel, H. J. *et al.* Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. *Nat. Biotechnol.* **25,** (2007).
9. Andersen, M. R. *et al.* Comparative genomics of citric-acid-producing *Aspergillus niger* ATCC 1015 versus enzyme-producing CBS 513.88. *Genome Res.* **21,** 885–97 (2011).
10. Wösten, H., Scoltmeijer, K. & de Vries, R. in *Food Mycology: A multifaceted approach to fungi and food* (eds. Dijksterhuis, J. & Samson, R.) 183–196 (2007).
11. Meijer, M., Houbraken, J. A. M. P., Dalhuijsen, S., Samson, R. A. & de Vries, R. P. Growth and hydrolase profiles can be used as characteristics to distinguish *Aspergillus niger* and other black aspergilli. *Stud. Mycol.* **69,** 19–30 (2011).
12. Association of Manufacturers and Formulators of Enzyme Products. List of commercial enzymes. (2009).

13. Workman, M., Andersen, M. R. & Thykaer, J. Integrated Approaches for Assessment of Cellular Performance in Industrially Relevant Filamentous Fungi. *Ind. Biotechnol.* **9,** 337–344 (2013).
14. Hong, S.-B. *et al. Aspergillus luchuensis*, an industrially important black *Aspergillus* in East Asia. *PLoS One* **8,** e63769 (2013).
15. Perrone, G. *et al. Aspergillus niger* contains the cryptic phylogenetic species *A. awamori*. *Fungal Biol.* **115,** 1138–1150 (2011).
16. Futagami, T. *et al.* Genome Sequence of the White Koji Mold *Aspergillus kawachii* IFO 4308, Used for Brewing the Japanese Distilled Spirit Shochu. *Eukaryot. Cell* **10,** 1586–1587 (2011).
17. Abarca, M. L., Bragulat, M. R., Castella, G. & Cabanes, F. J. Ochratoxin A production by strains of *Aspergillus niger* var. *niger*. *Applied and Environmental Microbiology* **60,** 2650–2652 (1994).
18. Frisvad, J. C., Smedsgaard, J., Samson, R. A., Larsen, T. O. & Thrane, U. Fumonisin B2 production by *Aspergillus niger*. *J. Agric. Food Chem.* **55,** 9727–32 (2007).
19. Frisvad, J. C. *et al.* Fumonisin and ochratoxin production in industrial *Aspergillus niger* strains. *PLoS One* **6,** (2011).
20. Perrone, G. *et al.* Biodiversity of *Aspergillus* species in some important agricultural products. in *Studies in Mycology* **59,** 53–66 (2007).
21. Monod, M. *et al.* Secreted proteases from pathogenic fungi. *Int. J. Med. Microbiol.* **292,** 405–19 (2002).
22. Jurjević, Z. *et al.* Two novel species of *Aspergillus* section Nigri from indoor air. *IMA Fungus* **3,** 159–73 (2012).
23. Varga, J. *et al.* New and revisited species in *Aspergillus* section Nigri. *Stud. Mycol.* **69,** 1–17 (2011).
24. Samson, R. A., Houbraken, J. A. M. P., Kuijpers, A. F. A., Frank, J. M. & Frisvad, J. C. New ochratoxin A or sclerotium producing species in *Aspergillus* section Nigri. *Stud. Mycol.* **50,** 45–61 (2004).
25. Samson, R. A. *et al.* Diagnostic tools to identify black aspergilli. *Stud. Mycol.* **59,** 129–45 (2007).
26. Samson, R. A. *et al.* Phylogeny, identification and nomenclature of the genus *Aspergillus*. *Stud. Mycol.* **78,** 141–73 (2014).
27. Visagie, C. M. *et al. Aspergillus*, *Penicillium* and *Talaromyces* isolated from house dust samples collected around the world. *Stud. Mycol.* **78,** 63–139 (2014).
28. Rajendran, C. & Muthappa, B. N. *Saitoa*, a new genus of *Plectomycetes*. *Proc. Plant Sci.* **89,** 185–191
29. Frisvad, J. C. *et al.* Formation of Sclerotia and Production of Indoloterpenes by *Aspergillus niger* and Other Species in Section Nigri. *PLoS One* **9,** e94857 (2014).
30. Nielsen, K. F., Mogensen, J. M., Johansen, M., Larsen, T. O. & Frisvad, J. C. Review of secondary metabolites and mycotoxins from the *Aspergillus niger* group. *Analytical and Bioanalytical Chemistry* **395,** 1225–1242 (2009).
31. Pel, H. J. *et al.* Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* {CBS} 513.88. *Nat. Biotechnol.* **25,** (2007).
32. Andersen, M. R. *et al.* Comparative genomics of citric-acid-producing *Aspergillus niger* ATCC 1015 versus enzyme-producing CBS 513.88. *Genome Res.* **21,** 885–897 (2011).
33. Yamada, O. *et al.* Genome sequence of *Aspergillus luchuensis* NBRC 4314. *DNA Res.* (2016). doi:10.1093/dnares/dsw032
34. de Vries, R. P. *et al.* Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus *Aspergillus*. *Genome Biol.* **18,** 28 (2017).
35. Kozakiewicz, Z. *et al.* Proposals for nomina specifica conservanda and rijicienda in *Aspergillus* and *Penicillium* (Fungi). *Taxon* **41,** 109–113 (1992).
36. Grigoriev, I. V. *et al.* MycoCosm portal: Gearing up for 1000 fungal genomes. *Nucleic Acids Res.* **42,** 699–704 (2014).
37. Grigoriev, I. V., Martinez, D. A. & Salamov, A. A. Fungal genomic annotation. *Appl. Mycol. Biotechnol.* **6,** 123–142 (2006).
38. Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37,** (2008).

39. Mitchell, A. *et al.* The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* **43,** D213–D221 (2015).
40. Sørensen, A., Lübeck, P. S., Lübeck, M., Teller, P. J. & Ahring, B. K. β-glucosidases from a new *Aspergillus* species can substitute commercial β-glucosidases for saccharification of lignocellulosic biomass. *Can. J. Microbiol.* **57,** 638–650 (2011).
41. Sørensen, A. *et al.* Identifying and characterizing the most significant β-glucosidase of the novel species *Aspergillus* saccharolyticus. *Can. J. Microbiol.* **58,** 1035–1046 (2012).
42. Karaffa, L. & Kubicek, C. P. *Aspergillus niger* citric acid accumulation: do we understand this well working black box? *Appl. Microbiol. Biotechnol.* **61,** 189–96 (2003).
43. Andersen, M. R., Nielsen, M. L. & Nielsen, J. Metabolic model integration of the bibliome, genome, metabolome and reactome of *Aspergillus niger*. *Mol. Syst. Biol.* **4,** 178 (2008).
44. Hossain, A. H. *et al.* Rewiring a secondary metabolite pathway towards itaconic acid production in *Aspergillus niger*. *Microb. Cell Fact.* **15,** 130 (2016).
45. Benoit, I. *et al.* Closely related fungi employ diverse enzymatic strategies to degrade plant biomass. *Biotechnol. Biofuels* **8,** 107 (2015).
46. Fox, E. M. & Howlett, B. J. Secondary metabolism: regulation and role in fungal biology. *Curr. Opin. Microbiol.* **11,** (2008).
47. Nielsen, J. C. *et al.* Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in *Penicillium* species. *Nat. Microbiol.* **2,** 17044 (2017).
48. Ali, H. *et al.* A non-canonical NRPS is involved in the synthesis of fungisporin and related hydrophobic cyclic tetrapeptides in *Penicillium chrysogenum*. *PLoS One* **9,** (2014).
49. Andersen, M. R. *et al.* Accurate prediction of secondary metabolite gene clusters in filamentous fungi. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E99-107 (2013).
50. Frandsen, R. J. N. *et al.* The biosynthetic pathway for aurofusarin in *Fusarium graminearum* reveals a close link between the naphthoquinones and naphthopyrones. *Mol. Microbiol.* **61,** (2006).
51. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).
52. Katoh, K. & Standley, D. M. MAFFT: multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
53. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56,** 564–577 (2007).

Chapter 3

Uncovering bioactive compounds in *Aspergillus* section *Nigri* by genetic dereplication using secondary metabolite gene cluster networks

The vast amount of bioactive compounds produced by Aspergilli makes them a great resource of new pharmaceuticals. Secondary metabolite gene clusters (SMGCs) are the genetic basis for these compounds and are also an defining the ecology of fungi. In our effort to describe the SMGCs of the whole *Aspergillus* genus, we redefine genome mining by exploiting homologous genes through species. We create similarity networks of SMGCs to sort them into non-redundant families. Our analysis reveals the genetic basis for analogous SM pathways throughout *Aspergillus* section Nigri and determines shared SMGC families between species on isolate, clade, section and genus level, illustrating their dynamics and evolution. Another application is to characterize unknown SMGC through guilt-by-association with known examples inside their family — a process we termed genetic dereplication. Genetic dereplication facilitated genome mining and target prioritization in newly sequenced strains. The final extension of our approach elucidates the genes for unlinked SMs with valuable bioactivities. Here, we correlate the presence of SMGC families through species and information on producing species to predict a non-ribosomal peptide synthetase necessary for production of the anticancer compound enhancing malformins (Wang et al., 2015b) in 18 strains. To illustrate the general applicability of our analyses, we decided to establish genetic engineering tools in A. brasiliensis; a species which has not previously been engineered nor used for elucidation of SMGC. We verified the predicted malformin synthetase gene *mlfA* by creating an A. brasiliensis *mlfA* deletion strain. Subsequent complementation of the knock-out strain with *mlfA* restored malformin production, demonstrating the accuracy of our prediction, and identifying the genetic basis of this pharmaceutically interesting compound.

Uncovering bioactive compounds in *Aspergillus* section *Nigri* by genetic dereplication using secondary metabolite gene cluster networks

Sebastian Theobald[a], Tammi Vesth[a], Jakob Kræmmer Rendsvig[a], Kristian Fog Nielsen[a], Robert Riley[b], Lucas Magalhães de Abreu[c], Asaf Salamov[b], Jens Christian Frisvad[a], Thomas Ostenfeld Larsen[a], Mikael Rørdam Andersen[a,1], and Jakob Blæsbjerg Hoof[a,1]

[a]Department of Biotechnology and Biomedicine, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark; [b]Department of Energy Joint Genome Institute, Walnut Creek, CA, USA; [c]Department of Plant Pathology, Federal University of Viçosa, Viçosa, Brasil

This manuscript was compiled on February 1, 2018

The increased interest in secondary metabolites (SMs) from fungi has driven the number of genome sequencing projects to elucidate their biosynthetic pathways. As a result, studies revealed that the number of secondary metabolite gene clusters (SMGCs) greatly outnumbers detected compounds, challenging current methods to dereplicate and categorize this amount of gene clusters on a larger scale. Here, we present an automated workflow for the genetic dereplication and analysis of secondary metabolism genes in fungi. Focusing on the secondary metabolite rich genus *Aspergillus*, we categorize SMGCs across genomes into SMGC families using network analysis. Our method elucidates the diversity and dynamics of secondary metabolism in section *Nigri*, showing that SMGC diversity within the section has the same magnitude as within the genus. Furthermore, we show how these families can be used to mine for gene clusters responsible for the production of specific compounds. Using our genome analysis we were able to predict the gene cluster of the potential anti-cancer compound malformin in 18 strains. To proof the general validity of our predictions, we developed genetic engineering tools in *Aspergillus brasiliensis* and subsequently verified the genes for biosynthesis of malformin.

Aspergillus | comparative genomics | genetic dereplication | natural products | malformin

The genus *Aspergillus* is one of the best studied fungal genera, with important species in the industrial, food and medical sector as well as in basic research. Its diverse repertoire of bioactive SMs e.g. anti-cancer compound enhancing malformins, cholesterol-lowering statins, and the mycotoxin aflatoxin have been detected in numerous analytical studies (6) — with many SMs applied primarily in the medical industry (7).

SMs are synthesized by different classes of enzymes. In fungi, these are polyketide synthases (PKS), non-ribosomal peptide synthetases (NRPS), terpene cyclases (TC), dimethylallyl transferases (DMATS), enzymes only consisting of a smaller subset of modules (PKS-Likes, NRPS-Likes), and fusions of PKS and NRPS (PKS-NRPS/ NRPS-PKS hybrids). These enzymes produce a SM backbone which is further modified by tailoring enzymes. The collective of enzymes necessary for production of a SM is encoded by a gene cluster.

NRPSs constitute a major group of secondary metabolite enzymes and can utilize L-amino acids, as well as non-proteogenic amino acids as their substrate (8), creating a diverse portfolio of compounds. Domains inside NRPSs are adenylation domains (A) for loading of amino acids, thiolation (T) domains for peptide chain transfer, condensation domains (C) for peptide bond formation, and epimerisation domains (E) to change the chirality of their proximate amino acid. Most NRPSs investigated show a colinearity rule, meaning they are assembled as modules in the order ATC. *Euascomycete* specific groups of NRPSs show substantial gain and loss of domains, further emphasizing the role of this enzyme class in chemical evolution of fungi (11). Understanding these dynamics and describing the diversity of NRPSs (and other secondary metabolites) throughout the genus *Aspergillus* will lead the way for new pharmaceutical drugs.

Prediction pipelines such as SMURF (12) and antiSMASH (13) facilitated the mining of genomic sequences for SMGC. Studies using these algorithms on fungal genomes revealed an amount of SMGCs exceeding expectations based on chemical analysis by far (14) — suggesting the biosynthetic potential of fungi is larger than anticipated. To efficiently analyse these large datasets over several organisms, genome neighbourhood networks have been used previously in bacteria to predict new gene clusters and ease strain prioritization for polyketides of interest (15). In this study, we used comparative genomics and network analysis to describe the dynamics and diversity of annotated and non-annotated SMGCs of 26 species of the SM-rich *Aspergillus* section *Nigri* (6) and five reference species. Identifying homologous gene clusters on the isolate, clade and section level enabled us to define groups with similar SMGC

Significance Statement

Fungi harbor a vast amount of secondary metabolites that are of great interest for society. Linking genes to metabolites, however, is a laborious task. Analyzing multiple genome sequences and comparing them on a genus level enables us to identify genes responsible for these metabolites, and describe evolutionary patterns and dynamics. Using a network approach, we make the algorithm reliable and scalable to more species and genomes, allowing fast expansion of the analysis. As an example, we link metabolic data to evolutionary patterns and directly identify the synthetase responsible for the anticancer agent malformin.

S.T., M.R.A., T.V., J.B.N., T.O.L., designed research; S.T., M.R.A., T.V., J.K.R., J.B.N., K.F.N., R.R., A.S., L.A. performed research; S.T., T.V., J.B.N., analyzed data; and S.T., M.R.A., J.B.N., J.K.R. wrote the paper.

The authors declare no conflict of interest.

[1]M.R.A. (E-mail: mr@bio.dtu.dk) is corresponding author on the computational analysis and J.B.H. is corresponding author on the experimental work (jblni@dtu.dk)

content inside section *Nigri*.

As an extension of our approach, we demonstrate the use of SMGC families together with information on metabolite profiles to mine for the gene cluster responsible for malformin biosynthesis. Malformins, a major group of compounds abundant in section *Nigri* (16), show anti-tobacco mosaic virus activity (17) and act as potentiator of anti-cancer drugs in mouse and human colon carcinoma cells (18). Identifying the SMGC responsible for malformin biosynthesis will allow for optimization of native as well as heterologous gene expression. Our approach successfully predicts the malformin gene cluster in multiple *Aspergillus* species, unveiling the feasibility of performing large scale dereplication of homologous gene clusters using collections of genome-sequenced strains.

Results

Creating families of secondary metabolite gene clusters. In order to describe the secondary metabolite gene cluster diversity of section *Nigri*, we analyzed genomes of 32 Aspergilli plus the reference species *A. oryzae*, *A. flavus*, *A. fumigatus*, *A. nidulans* and *Penicillium chrysogenum*. Aspergilli are known to produce similar SMs throughout species (6, 19, 20), thus including these species would ensure to relate SMGC content to phylogeny and potentially reveal homologous gene clusters.

All 37 genome sequences were analyzed using our previously established pipeline (21). The pipeline creates families of homologous gene clusters using local alignment of protein sequences of cluster members. It retains bidirectional hits that are above the coverage cutoff and uses the percent identity to compute a similarity score for each query cluster to each hit cluster. Subsequently, the similarity scores are used to create a network of all SMGCs and random walk clustering is used to create families of SMGCs. Using the pipeline on the dataset, we detected 2,717 gene clusters and categorized them into 455 families — groups predicted to produce similar compounds based on homologous gene clusters — of which 245 only contain one cluster and are therefore unique (21).

Comparative genomics reveals secondary metabolite gene cluster diversity on several taxonomic levels in section *Nigri*. With the establishment of SMGC families in related species, we were interested whether SMGC content is reflected in species, clade, section and genus circumscription. Differences in SM content have been shown previously for the groups of uniseriates (species with phialides attached directly to the vesicle) and biseriates (species with metulae between phialides and vesicle) inside section *Nigri* (Fig. 1, (6)). Hence, differences in SM production should be the result of different SMGCs present in species.

Comparing all shared SMGCs throughout species in the dataset, we established a dendrogram based on the similarities of SMGCs between fungi (Fig. 1), which to a large extent reflects the phylogenetic tree of Aspergilli (left side of Fig. 1 panel A ; (21)) with subtle differences. The heatmap in figure 1 provides further insight into SMGC similarity of species. Clustering the organisms by their SMGC families shows a clear distinction between biseriate species, uniseriate species and reference species. Inside these groups, further subgroups can be identified.

A. niger isolates shared 80–100% of SMGC families — pointing out that isolates of the same species can carry a few different SMGCs — although none of them have unique SMGCs. The shared SMGC content among species varied depending on the clade. Species in the *A. niger* and *A. tubingensis* clade share 60–80% — with *A. eucalypticola* showing a distinct SMGC composition inside the clade. Species in the *A. carbonarius* clade share 60–70% of SMGC families. This similarity dropped to 50–60% inside the *A. heteromorphus* clade. Most uniseriates shared at least 60% SMGCs, with the exception of *A. saccharolyticus* and *A. homomorphus* only sharing as few as 40% of their SMGC families with other members of the uniseriates. On a section level, we can show that biseriates and uniseriates (apart from the *A. heteromorphus* clade) each show a SMGC family inter-clade similarity of at least 40%. Comparing the *A. tubingensis* and *A. niger* clade to uniseriates the SMGC similarity is 10–30% — the same as between the section *Nigri* and the reference species. Hence, we can determine that the diversity of secondary metabolites inside section *Nigri* is similar to the diversity seen across the genus as a whole.

From this, it can be inferred that section *Nigri* must have undergone a substantial gain and loss of secondary metabolite genes. In species which show a larger difference in SMGC composition to closely related species — as in the case of *A. eucalypticola* — suggests horizontal gene transfer from outside section *Nigri* or retention of SMGC. Surprisingly, a small amount of SMGCs seem to be retained in the whole genus since we find at least 10% similarity of SMGC families between the species included in the analysis. Additionally, a maximum of 30% shared SMGC families between distantly related species exceeds the SMGC similarity previously anticipated in the genus *Aspergillus* (22).

In conclusion, we can confirm that the clustering of SMGCs into families reflects the SM distribution of species in analytical studies. The SMGC similarity over large phylogenetic distances suggests analogous pathways in the same family.

Coupling MIBiG annotation to SMGC families automates genetic dereplication of compounds. With the diversity of SMGCs established through our dataset, we were interested in the presence of gene clusters linked to known compounds through section *Nigri*. With an increasing number of available fungal genome sequences, we see an identification of known compounds by genomic methods as crucial, since laboratory conditions might not reveal the full metabolite profile of a fungus. Furthermore, it may help to avoid experiments re-identifying the same gene cluster in multiple species (similar to the process known as metabolite dereplication in chemical analysis (23)). To achieve this, we used 1461 gene clusters of the Minimum Information about a Biosynthetic Gene cluster (MIBiG) database (24) to identify known compounds with characterized SMGCs in our SMGC families and determine related compounds. This is of special interest for mycotoxins and compounds with medical applications.

Using protein BLAST (25), we identified 34 best hits found in our SMGC families for compound-linked gene clusters retrieved from MIBiG. Since SMGC families represent groups of homologous and related gene clusters, we can identify the SMGC family of the hit as a related gene cluster producing a similar compound by using a guilt by association approach. Hence, new genomes can be analyzed and added with information on their secondary metabolite production capabilities. The associated compounds and presence patterns of gene clus-

ters are shown in Fig. 1 B.

Of the 34 known gene clusters used to annotate the SMGC families in the dataset, two gene clusters linked to the compounds fungisporin, YWA and one gene cluster linked to the siderophore ferrichrome were found in all species of the dataset (21). This illustrates that we can detect homologous gene clusters over the genus as similarly shown by the shared gene cluster overview (Fig.1). 14 gene clusters were only found in the reference species.

SMGCs for different heteroisoextrolites based on 6-MSA are found in section Nigri. Aspergili in distinct sections are known to produce functionally similar types of secondary metabolites, also called heteroisoextrolites (20). These heteroisoextrolites are based on analogous biosynthetic pathways which we successfully annotated in gene cluster families.

Using our automated method, we were able to detect SMGCs for heteroisoextrolites that are based on 6-methylsalicylic acid (6-MSA), in particular the antifungal patulin (2, 26) and the antimicrobial yanuthone D (3, 27). Association of the patulin gene cluster with its family identifies 11 patulin-like gene clusters in uniseriates. Further inspection of the family revealed the cluster responsible for aculinic acid in *A. aculeatus* which is shown to be highly similar in genetic content and function to the patulin gene cluster (28). Similarly, gene clusters primarily found in the *A. niger* clade, as well as in *Penicillium chrysogenum* are predicted to produce the antifungal 6-MSA-based compound yanuthone (27, 29). The network plot in Fig. S1 shows how the related clusters were divided into families and highlights how SMGC networks can be used to classify related SMGCs.

SMGCs related to the trypacidin gene cluster can be found in uniseriate species. Another group of heteroisoextrolites present throughout Aspergilli are compounds based on emodin, such as geodin and the mycotoxins trypacidin and secalonic acid (20). Secalonic acid is of special interest due to a wide range of bioactivities (1, 4). We identified a SMGC family containing the monodictyphenone/emodin gene cluster (30) from *A. nidulans*, the trypacidin gene cluster from *A. fumigatus* (31) and several unlinked gene clusters in uniseriates. Uniseriate species are known to produce secalonic acid (19, 32), hence we predicted the unlinked gene clusters in this family to be responsible for secalonic acid production. The presence in *A. brasiliensis* is surprising, but according to the analysis of shared SMGC (Fig. 1), it is likely that it acquired or retained SMGCs from other species and therefore stands out of the *A. niger* and *A. tubingensis* clade.

Genetic dereplication predicts a gene cluster for azanigerone-like compounds in 17 strains. Previous studies identified the silent azanigerone gene cluster by overexpression of cluster genes in *Aspergillus* niger ATCC 1015 (33). In our analysis, we can identify an azanigerone-like producing gene cluster in biseriates, and *A. homomorphus* — which is uniseriate (supplementary file 1). This further highlights our algorithm as an important addition to chemical analysis, since genetic dereplication is able to identify gene clusters over a large set of genomic sequences, even though they may be silent in the hosts under normal conditions.

SMGC families contain communities of related SMGC. Gene clusters producing the structurally related polyketides asperthecin and TAN-1612 (a neuropeptide Y antagonist (34)) can be found in the same SMGC family. The gene cluster in *A. nidulans* is producing asperthecin, while the gene clusters predicted in the *A. niger* and *A. tubingensis* clades are likely producing TAN-1612 since the uniform presence in these two clades suggests the gene cluster to be conserved throughout species (Fig. 1 B, supplementary file 1). This highlights further that our method can be used to mine for similar compounds in SMGC families.

A gene cluster predicted in the *A. niger* clade and *A. tubingensis* is highly similar to the pseurotin A gene cluster, a chitin synthase inhibitor (35) from *A. fumigatus* (36), where it forms a super cluster together with fumagillin (37).

Uniseriates contain a related gene cluster to *A. flavus* (39) and *A. oryzae* (40) responsible for the production of food contaminant and mycotoxin cyclopiazonic acid (CPA). This further confirms our findings that a number of SMGCs can be shared over large phylogenetic distances (Fig. 1). It is surprising that the gene cluster is also found in *A. saccharolyticus* and *A. heteromorphus* since they differ in their SMGC content from the rather SM homogeneous rest of uniseriates (Fig. 1). Kato et al. (41) highlighted that CPA is produced but converted in *A. oryzae*, so it remains to be answered if the uniseriate species produce CPA or a derivate thereof. A gene cluster unique to uniseriates is predicted to be responsible for aculeacin A production. Aculeacin A is an interesting lipopeptide inhibiting β-glucan synthesis (42).

All in all, our analysis can automatically identify related SMGCs over a large set of species. However, our analysis also highlights that genetically dereplicated SMGC only constitute a small fraction of the secondary metabolites potentially produced by Aspergilli.

Mining for gene clusters in SMGC families reveals candidates for the malformin gene cluster in 18 strains. To address the large interest in discovery of novel biosynthetic gene clusters for specific compounds, we wanted to link SMGC families to compounds of interest. For this, we focused on malformin-producing species: *A. niger*, *A. brasiliensis*, and *A. tubingensis* (6). Malformin is interesting as it is both a potential compound for medical application in cancer treatment (18) and is produced under laboratory conditions.

The first criterium for a malformin-producing gene cluster was set to a gene cluster being present in all producing species. Since the species mentioned above share 50% or more of their SMGC families, more levels of filtering are required to narrow down candidates. The second criterium considered the size of candidate NRPSs. Malformin is a pentapeptide consisting of Val, D-Leu, Ile and two D-Cys amino acids. Hence, it was anticipated the NRPS would consist of five modules with a length of approximately 18kb. As we could not find a hit with these parameters, we moderated our search to four modules under the hypothesis that one of the modules is iterative (as seen in other studies (43)), resulting in a size of approximately 12kb. The third criterium considered the annotation of tailoring enzymes present within the cluster. Tailoring enzymes should contain disulphide bond associated enzymes to be able to create the disulphide bond included in malformins (Fig. 4 D).

Using our search criteria, we found a single candidate NRPS gene cluster matching all of our criteria. The SMGC family shows a high level of synteny and all gene clusters inside the family show the same NRPS size and disulphide bond creating

enzymes (Fig. 2). Inside the cluster, we found genes coding for major facility superfamily transporters, methyltransferases and a transcription factor — activities common to SMGCs. The gene cluster family was curated by removing four gene clusters shown in Fig. S2, which only aligned to the extended part of the predicted cluster and do not contribute to malformin synthesis.

Predicting condensation domain function in MlfA. To further confirm the candidate cluster, we created a maximum likelihood phylogeny of NRPS condensation domains with known functions in our dataset (fungisporin/ nidulanin A (5, 44, 45), fumiquinazolines (46), fumitremorgin/ brevianamide (36) and penicillin (47)), including condensation domains of the predicted malformin synthetase MlfA to predict their functions (Fig. 3).

According to the amino acid composition of malformin, we expected epimerization, epimerizing D-L joining condensation domains and a cyclizing condensation domain to be present in the synthetase. From branches in the phylogeny, we can predict the functions of the five condensation domains in malformin to be DL-joining (epimerizing subtype), LL-joining, epimerization, DL-joining and cyclizing domain, thus matching the expectations and supporting the identification of the NRPS as involved in malformin production.

Genetic and chemical analysis verifies *mlfA* prediction. To verify the genetic assignment, we first had to develop genetic engineering tools in *A. brasiliensis*. We decided to construct a *pyrG*Δ strain in order to subsequently generate a non-homologous end-joining deficient strain — facilitating efficient gene targeting (48). We employed a clustered regularly interspaced short palindromic repeats associated endonuclease 9 (CRISPR-Cas9) system (49) to induce a double-strand break (DSB) in *pyrG* resulting in uridine auxotrophy. Subsequent sequencing of three candidates confirmed that strain 1 had an out-of-frame mutation in the region corresponding to the protospacer via a 16-nucleotide deletion (nucleotides number 45-60, allele name *pyrG1*) within *pyrG*. In this strain, we utilized the CRISPR-Cas9 system to induce a DSB at the *akuA* locus while supplying a repair template in form of a linear gene-targeting substrate for *akuA*. A correct homokaryotic transformant was verified as an *akuA* deletant by diagnostic tissue polymerase chain reaction (PCR) (Fig. S3). The strain, *akuA*Δ::AFL*pyrG*, was screened on 5-fluoroorotic acid (5-FOA) enriched growth medium for loss of AFL*pyrG* by the lack of ability to grow without supplementation of uridine in the medium and diagnostic PCR. In the resulting *pyrG*-free strain, *akuA*Δ, we targeted the NRPS encoded by Aspbr1_34020, which based on the predictions above was the best candidate for malformin production. Six transformants were streak-purified and PCR analyzed, resulting in two homokaryotic deletion mutants of *mlfA* (see Fig. S3). Both strains were subsequently screened, alongside *akuA*Δ::AFL*pyrG* as reference, for their ability to produce malformin A and C after seven days of cultivation on yeast extract sucrose (YES) solid growth medium. The deletion of Aspbr1_34020 (*mlfA*Δ) showed a total abolishment of malformin production (Fig. 4 C).

Moreover, a genetic complementation by a constitutively expressed *mlfA* in the *mlfA*Δ strain re-established malformin production with the same adduct pattern as for the reference strain, thus confirming the role of *mlfA* in malformin production (Fig. 4 C).

Discussion

With the first *Aspergillus* genome sequences available (51, 52), research on natural products and on the evolution of secondary metabolism in fungi experienced a paradigm shift (53, 54). Now with whole genus sequencing projects, these fields are experiencing a second shift due to the amount of generated data. This study fills the gap for automated and scalable multi-species dereplication and classification of secondary metabolite gene clusters using similarity networks. Specific to our study is the mapping of phylogeny on SMGC families which then enables a relation of phylogeny to SMGC content. Hence, we can specifically determine the amount of SMGC shared between certain species.

Using SMGC similarity networks and random walk clustering we grouped SMGCs across 37 genomes into families of gene clusters. Remarkably, we found that species on the isolate level can carry a few distinct SMGC as shown by a similarity of 80-90% between *A. niger* isolates. Previous studies have shown that the *A. niger* group is prone to undergo extensive transfer and loss of genetic elements, also reflected in production of different exometabolites (55). Species of distinct clades inside section *Nigri* can share 30–80 % depending on the distance of the species, showing a diversity similar to *Penicillia* clades as indicated recently (56). Gene cluster families over different clades are of special interest since they harbour similar compounds as heteroisoextrolites (20). Examples in this study include SMGC based on emodin as e.g. trypacidin, secalonic acid and SMGC based on 6-MSA that we could link to yanuthone-D, aculinic acid and patulin throughout different species of the dataset. 6-MSA based heteroisoextrolites seem to be common in *Trichocomaceae* since Penicillia contain yanuthone producing gene clusters as well (56). The most phylogenetically distinct clades exhibited a surprising amount of diversity in their SMGC content. In comparisons between uniseriate and biseriate species, we could show a SMGC similarity of 10-20%. We gained similar results in comparisons of *Nigri* species with reference species. Thus, the SMGC diversity within the section has the same magnitude as within the whole genus *Aspergillus*. This finding is in accordance with analytical studies characterizing biseriates and uniseriates as distinct groups based on their metabolite production (6). On a genus level, the amount of shared SMGC families over large phylogenetic distances is higher in our study than estimated by Lind et al. (22).

As a result of our analysis, we were able to predict the malformin gene cluster in 18 strains and confirmed it in *A. brasiliensis*. Our results are in accordance with reports of producing strains as mentioned by Nielsen et al. and Vesth et al. (6, 21). Surprisingly, the predicted NRPS consists of only four modules which did not match our assumptions of an NRPS structure for five amino acids. Examples of NRPS following different synthesis schemes, however, are increasing. Enniatin, enterobactin, bacillibactin and gramicidin S are synthesized by iterative NRPSs (9). Identification of tailoring enzymes coding for disulphide-bond associated functions and establishment of a condensation domain model for the predicted gene cluster/ synthetase helped us to further sustain our prediction. A phylogenetic characterization of condensation domains was chosen over prediction tools since these tools were built on bacterial

data (57). Upon deletion of *mlfA*, malformin production is abolished. Furthermore, we were able to show that complementation of *mlfA*Δ strains with *mlfA* can revive production of malformins. In combination, this makes us confident that *mlfA* is coding for a NRPS responsible for malformin production. We hypothesize the NRPS to act iteratively on one amino acid and possesses multiple amino acid specificities since multiple malformins disappear after deletion of the NRPS (Fig. 4).

Our study shows that SMGC similarity networks and families are ideal constructs for guilt by association based dereplication and genome mining for SMs of interest. This also shows the advantage of using an aggregated similarity score per gene cluster as opposed to one comparison per domain, since only one SMGC will be associated with a compound (56). Another advantage of our algorithm is to genetically dereplicate SMGCs which would go unnoticed with existing methods. We were able to identify homologs of a gene cluster in 17 strains, which is silent in the original host under laboratory conditions (33). Hence, genetic dereplication uncovers new sets of SMGCs as targets for heterologous expression that might not be discovered by, e.g. OSMAC (58). Our study also shows that the amount of dereplicated SMGCs in section *Nigri* is far lower than the total amount of SMGCs (21). Hence this study will facilitate further efforts to investigate the SMGCs of section *Nigri*.

Genome sequencing projects are extending in their amount of data and are in need of automated methods for dereplication and characterization (14, 21, 56, 59, 60). This study presents a pipeline that will cope with the challenges of increasing genomic datasets and pinpoint towards related SMGC for synthetic biology approaches. With the relational characterization of a whole section of species at once, our study will serve as a resource for future analyses of SMGCs and SMs in Aspergilli. Especially similarity networks and guilt by association dereplication of genome data will facilitate genome mining efforts.

Finally, similarity networks of SMGCs prove to serve for the genetic dereplication of SMGCs in several species and establish their phylogenetic distribution. Assessing and categorizing the metabolic potential of species in this automated manner will greatly facilitate the discovery of new relevant SMs.

Materials and Methods

Data collection. A customized version of SMURF (12) was used to annotate secondary metabolite gene clusters throughout draft *Aspergillus* genomes. Protein sequences, smurf annotations, interpro annotations and gff files were obtained from jgi (https://genome.jgi.doe.gov/).

Draft genomes of 32 Aspergilli of section *Nigri* and five other *Aspergillus* genomes were analyzed using our previously established pipeline to study their secondary metabolite gene cluster diversity (21).

Creation of SMGC families. Families of gene clusters were created with the pipeline established in (21). BLAST+ (25) was used to find bidirectional hits between all secondary metabolite proteins using an E-value of 1e-10, at least 50% identity and a sum of coverage of 130% as cutoffs. Subsequently, the identity values were aggregated from query to hit clusters to create a cluster similarity score (21). The established connections were then used to create a network of secondary metabolite gene cluster proteins (61) and random walk clustering (62) in R (63) with 1 step was used to find families of homologous/ related gene cluster proteins. In case a family carried more members than available organisms the clustering was repeated with the family to find further subcommunities.

Visualization of shared SMGC families. A heatmap containing hirarchically clustered column dendrograms was created using the gplots package (64) and a phylogenetic tree imposed on rows (65).

Mining for malformin producing NRPS. Created SMGC families were classified as potential producers of malformin according to three criteria. Strains which are known to produce malformin (*A. niger* CBS 513.88, *A. niger* NRRL3, *A. niger* ATCC 1015, *A. brasiliensis*, and *A. tubingensis*) should be included in the family; the clusters should include an NRPS of at least 14400 nucleotides and tailoring enzymes should include the terms 'glutathione' or 'disulphide'. From the two resulting families, the best hit was used for further investigation. Gene clusters were visualized using Gviz (66).

Whole genome phylogeny. A whole genome phylogenetic tree was generated to compare phylogeny to hirarchical clustering based on secondary metabolite family content. The phylogeny was constructed using 200 bidirectional best hits between species. These best hits were concatenated and aligned using MAFFT (67) and conserved blocks extracted using Gblocks (68). A maximum likelihood phylogeny was created using the trimmed alignments for multithreaded RAxML with PROTGAMMAWAG model and 100 bootstraps (69).

Prediction of condensation domain types. Condensation domains were extracted from protein sequences using annotations from InterproScan5 (70). Separated domains, i.e. domains smaller than 350 amino acids and less than 100 amino acids apart from a domain of the same type, were merged before proceeding. Resulting domain sequences were aligned using Clustal Omega (71) and trimmed using trimal (72) retaining sequences with over 65 % residue coverage in over 80 % of sequences and removing all columns with gaps in more than 20 % of sequences with similarity lower than 0.001 but preserving at least 60 % of columns. IQ-tree (73) was used on aligned sequences using a LG+F+I+G4 substitution model (74) and 1000 times bootstrap (75). Functions of condensation domains of the predicted malformin NRPS could be assigned by coclustering with known examples.

Annotating SMGC families using MIBiG. Gene cluster annotations were downloaded from the MIBiG database (24) and 1461 sequences of backbone proteins extracted using biopython (38). Protein sequences were then blasted against our dataset. Hits reaching a percent identity, query coverage and hit coverage of over 95% were retained to find best hits in our dataset. Corresponding SMGC families were annotated as related cluster of the hit.

Construction of mutant strains. The wild type culture (WT) *A. brasiliensis* (CBS 101740/IBT 21946) (76) was used, to generate a uridine requiring pyrG- strain (pyrG1, BRA6), and from BRA6, a knockout strain of the Ku70 homolog akuA was created to enable efficient gene targeting (48), see Table S1 for strains. Genomic DNA (gDNA) from WT A. brasiliensis was isolated via FastDNA SPIN Kit for Soil DNA extraction kit (MP Biomedicals, USA).

All *A. brasiliensis* strains were cultivated at 30°C on minimal medium (MM), supplemented with 10 mM uridine if required for growth. The MM, transformation media (TM) and media for pyrG counter-selection (MM+5-FOA) were prepared as described in (49). All transformations employing CRISPR/Cas9 vectors used hygromycin B (100 μg/ml, Invivogen) for selection. Yeast extract sucrose (YES, (77)) growth media was used for chemical analysis. Chemical competent *Escherichia coli* DH5α were applied for vector assembly and plasmid propagation at 37°C, and *E. coli* cultivations were carried out in Lumia Broth (LB) media (1% Bacto tryptone, 0.5% Bacto yeast extract, 1% NaCl, pH 7.0) supplemented with 0.1% ampicillin. All solid media were supplied with 2% agar.

Secondary metabolite extraction and analysis. Extraction of secondary metabolites from solid media (CYA and YES) 6 plugs (6 mm) were based on samples across the radius of the fungal colony, transferred to a microcentrifuge tube and covered in ethyl acetate/2-propanol 3:1(v/v) with 1% (v/v) formic acid for 60 min

ultrasonication. The extraction solvent was transferred to a clean vial, solvents evaporated using N_2 flow, and the residues on the tube walls were re-dissolved in methanol for 30 min by ultrasonication. The samples were centrifuged at 15000 g and the supernatant transferred to a HPLC auto sampler vial. UHPLC-DAD-QTOFMS was performed on an Agilent Infinity 1290 UHPLC system equipped with a diode array detector. Separation was done on a 250 × 2.1 mm i.d., 2.7 μm, Poroshell 120 Phenyl Hexyl column (Agilent Technologies, Santa Clara, CA) held at 60°C. Subsamples of 1 μL, were eluted with a flow rate of 0.35 mL/min using A: water with 20 mM formic acid and B: acetonitrile with 20 mM formic acid as a gradient system starting at 90% A, which linearly dropped to 10% in 15 min, and held for 2 min before returning to 90% for 2 min. Acetonitrile, methanol, ethyl acetate, 2-propanol and formic acid were analytical grade (Sigma-Aldrich, St. Louis, MO, USA). Water, acetonitrile and formic acid for MS solvents were all LC-MS grade (Sigma-Aldrich). Mass spectrometry (MS) detection was performed on an Agilent 6545 QTOF MS equipped with an Agilent dual jet stream ESI operated in ESI^+ mode, with MS spectra recorded as centroid data, at an m/z of 100 to 1,700, and auto MS/HRMS fragmentation was performed at three collision energies (10, 20, 40 eV), on the three most intense precursor peaks per cycle. The acquisition was 10 spectra/s. Data were treated in Agilent MassHunter Qualitative Analysis, and compounds were detected using extracted ion chromatograms (EICs) ± m/z 0.005 Da of the theoretical masses (78). MSHRMS were evaluated against a database of 1500 compounds, while HRMS and MS/HRMS peaks were matched against around 3000 known and suspected *Aspergillus* compounds. Reference standards of malformins C and A were co-analysed in the sequence. Malformin A2 ($C_{22}H_{37}O_5N_5S_2$) and C ($C_{23}H_{39}O_5N_5S_2$) were detected at using EICs of expected adducts ($[M^+H]^+$, $[M^+NH_4]^+$, $[M^+Na]^+$) based on the calculated monoisotopic mass [M], 515.2236 Da and 529.2393 Da.

ACKNOWLEDGMENTS. We thank Martin Engelhard Kogle and Ellen Kirstine Lyhne for gDNA preparation of *Aspergillus* strains. ST, TV, and MRA gratefully acknowledge support from the Villum Foundation, grant VKR023437. We thank Prof Adrian Tsang for making the *A. niger* NRRL3 genome available.

1. Hu et al.(2012Secalonic acid D reduced the percentage of side populations by down-regulating the expression of ABCG2 *Biochemical Pharmacology* 85(11):1619–1625.
2. Iwahashi et al.(2006)Mechanisms of patulin toxicity under conditions that inhibit yeast growth *Journal of Agricultural and Food Chemistry* 54(5):1936–1942.
3. Bugni et al. (2000)Yanuthones: Novel metabolites from a marine isolate of Aspergillus niger *Journal of Organic Chemistry* 65(21):7195–7200.
4. Zhai et al.(2013)Secalonic acid A protects dopaminergic neurons from 1-methyl-4-phenylpyridinium (MPP+)-induced cell death via the mitochondrial apoptotic pathway *European Journal of Pharmacology* 713(1–3):58–67.
5. Andersen et al.(2013)Accurate prediction of secondary metabolite gene clusters in filamentous fungi. *Proceedings of the National Academy of Sciences of the United States of America* 110(9):E99–E107.
6. Nielsen KF, Mogensen JM, Johansen M, Larsen TO, Frisvad JC (2009) Review of secondary metabolites and mycotoxins from the Aspergillus niger group. *Analytical and Bioanalytical Chemistry* 395(5):1225–1242.
7. Martínez-Núñez MA, et al. (2016) Nonribosomal peptides synthetases and their applications in industry. *Sustainable Chemical Processes* 4(1):13.
8. Finking R, Marahiel MA (2004) Biosynthesis of Nonribosomal Peptides. *Annual Review of Microbiology* 58(1):453–488.
9. Mootz HD, Schwarzer D, Marahiel MA (2002) Ways of assembling complex natural products on modular nonribosomal peptide synthetases. *ChemBioChem* 3(6):490–504.
10. Fischbach MA, Walsh CT, Clardy J (2008) The evolution of gene collectives: How natural selection drives chemical innovation. *Pnas* 105(12):4601–4608.
11. Bushley KE, Turgeon BG (2010) Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. *BMC evolutionary biology* 10(1):26.
12. Khaldi N, et al. (2010) SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal genetics and biology : FG & B* 47(9):736–41.
13. Medema MH, et al. (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic acids research* 39(Web Server issue):339–46.
14. de Vries RP, et al. (2016) *Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus Aspergillus.* pp. 1–45.
15. Rudolf JD, Yan X, Shen B (2016) Genome neighborhood network reveals insights into enediyne biosynthesis and facilitates prediction and prioritization for discovery. *Journal of Industrial Microbiology and Biotechnology* 43(2-3):261–276.
16. Yukioka M, Winnick T (1966) Synthesis of malformin by an enzyme preparation from Aspergillus niger. *Journal of Bacteriology* 91(6):2237–2244.
17. Tan QW, Gao FL, Wang FR, Chen QJ (2015) Anti-TMV activity of malformin A1, a cyclic pentapeptide produced by an endophytic fungus Aspergillus tubingensis FJBJ11. *International Journal of Molecular Sciences* 16(3):5750–5761.
18. Wang J, et al. (2015) Study of malformin C, a fungal source cyclic pentapeptide, as an anticancer drug. *PLoS ONE* 10(11):1–19.
19. Samson RA, et al. (2007) Diagnostic tools to identify black aspergilli. *Studies in mycology* 59:129–45.
20. Frisvad JC, Larsen TO (2015) Chemodiversity in the genus Aspergillus. *Applied Microbiology and Biotechnology* 99(19):7859–7877.
21. Vesth TC, et al. (2018) The genomes of Aspergillus section Nigri reveal drivers in fungal speciation (submitted). *Submitted to Nature.*
22. Lind AL, et al. (2015) Examining the Evolution of the Regulatory Circuit Controlling Secondary Metabolism and Development in the Fungal Genus Aspergillus. *PLOS Genetics* 11(3):e1005096.
23. Klitgaard A, et al. (2014) Aggressive dereplication using UHPLC-DAD-QTOF: screening extracts for up to 3000 fungal secondary metabolites. *Analytical and bioanalytical chemistry* 406(7):1933–43.
24. Medema MH, et al. (2015) Minimum Information about a Biosynthetic Gene cluster. *Nature Chemical Biology* 11(9):625–631.
25. Camacho C, et al. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10(1):421.
26. Tannous J, et al. (2014) Sequencing , physical organization and kinetic expression of the patulin biosynthetic gene cluster from Penicillium expansum. *International Journal of Food Microbiology* 189:51–60.
27. Petersen LM, et al. (2015) Characterization of four new antifungal yanuthones from Aspergillus niger. *The Journal of antibiotics* 68(3):201–5.
28. Petersen LM, et al. (2015) Investigation of a 6-MSA Synthase Gene Cluster in Aspergillus aculeatus Reveals 6-MSA-derived Aculinic Acid, Aculins A-B and Epi-Aculin A. *ChemBioChem* 16(15):2200–2204.
29. Holm DK, et al. (2014) Molecular and chemical characterization of the biosynthesis of the 6-MSA-derived meroterpenoid yanuthone D in Aspergillus niger. *Chemistry and Biology* 21(4):519–529.
30. Chiang YM, et al. (2010) Characterization of the aspergillus nidulans monodictyphenone gene cluster. *Applied and Environmental Microbiology* 76(7):2067–2074.
31. Mattern DJ, et al. (2015) Identification of the antiphagocytic trypacidin gene cluster in the human-pathogenic fungus Aspergillus fumigatus. *Applied microbiology and biotechnology* pp. 10151–10161.
32. Fungaro MHP, et al. (2017) Aspergillus labruscus sp. nov., a new species of Aspergillus section Nigri discovered in Brazil. *Scientific Reports* 7(1):1–9.
33. Zabala AO, Xu W, Chooi YH, Tang Y (2012) Discovery and Characterization of a Silent Gene Cluster that Produces Azaphilones from Aspergillus niger ATCC 1015 Reveal a Hydroxylation-Mediated Pyran-Ring Formation. *Chemistry & biology* 19(8):1049–59.
34. Kodukula K, et al. (1995) BMS-192548, a tetracyclic binding inhibitor of neuropeptide Y receptors, from Aspergillus niger WB2346. I. Taxonomy, fermentation, isolation and biological activity. *The Journal of antibiotics* 48(10):1055–9.
35. Wenke J, Anke H, Sterner (1993) Pseurotin A and 8-O-Demethylpseurotin A from Aspergillus fimugatus and theirs inhibitory activities osn chitin synthase. *Bioscience, Biotechnology, and Biochemistry* 57(6):961–964.
36. Maiya S, Grundmann A, Li SM, Turner G (2006) The fumitremorgin gene cluster of Aspergillus fumigatus: Identification of a gene encoding brevianamide F synthetase. *ChemBioChem* 7(7):1062–1069.
37. Wiemann P, et al. (2013) Prototype of an intertwined secondary-metabolite supercluster. *Proceedings of the National Academy of Sciences* 110(42):17065–17070.
38. Cock, P, et al. (2009)Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *11(25):1422–1423*
39. Varga J, Baranyi N, Chandrasekaran M, Vágvölgyi C, Kocsubé S (2015) Mycotoxin producers in the Aspergillus genus: An update. *Acta Biologica Szegediensis* 59(2):151–167.
40. Tokuoka M, et al. (2008) Identification of a novel polyketide synthase-nonribosomal peptide synthetase (PKS-NRPS) gene required for the biosynthesis of cyclopiazonic acid in Aspergillus oryzae. *Fungal Genetics and Biology* 45(12):1608–1615.
41. Kato N, et al. (2011) Genetic Safeguard against Mycotoxin Cyclopiazonic Acid Production in Aspergillus oryzae. *ChemBioChem* 12(9):1376–1382.
42. Hector RF (1993) Compounds active against cell walls of medically important fungi. *Clinical microbiology reviews* 6(1):1–21.
43. Juguet M, et al. (2009) An Iterative Nonribosomal Peptide Synthetase Assembles the Pyrrole-Amide Antibiotic Congocidine in Streptomyces ambofaciens. *Chemistry and Biology* 16(4):421–431.
44. Klitgaard A, Nielsen JB, Frandsen RJN, Andersen MR, Nielsen KF (2015) Combining Stable Isotope Labeling and Molecular Networking for Biosynthetic Pathway Characterization. *Analytical Chemistry* 87(13):6520–6526.
45. Ali H, et al. (2014) A non-canonical NRPS is involved in the synthesis of fungisporin and related hydrophobic cyclic tetrapeptides in Penicillium chrysogenum. *PLoS ONE* 9(6).
46. Gao X, et al. (2012) Cyclization of fungal nonribosomal peptides by a terminal condensation-like domain. *Nature chemical biology* 8(10):823–30.
47. Diez B, Ii V, Martin JF, Barredosll JL (1990) The Cluster of Penicillin Biosynthetic Genes. *Biochemistry* 265(27):16358–16365.
48. Nielsen JB, Nielsen ML, Mortensen UH (2008) Transient disruption of non-homologous end-joining facilitates targeted genome manipulations in the filamentous fungus Aspergillus nidulans. *Fungal genetics and biology : FG & B* 45(3):165–70.
49. Nødvig CS, Nielsen JB, Kogle ME, Mortensen UH (2015) A CRISPR-Cas9 system for genetic engineering of filamentous fungi. *PLoS ONE* 10(7):1–18.
50. Chung BKW, Yudin AK (2015) Disulfide-bridged peptide macrobicycles from nature. *Organic & Biomolecular Chemistry* 13(33):8768–8779.
51. Galagan JE, et al. (2005) Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. *Nature* 438(7071):1105–1115.

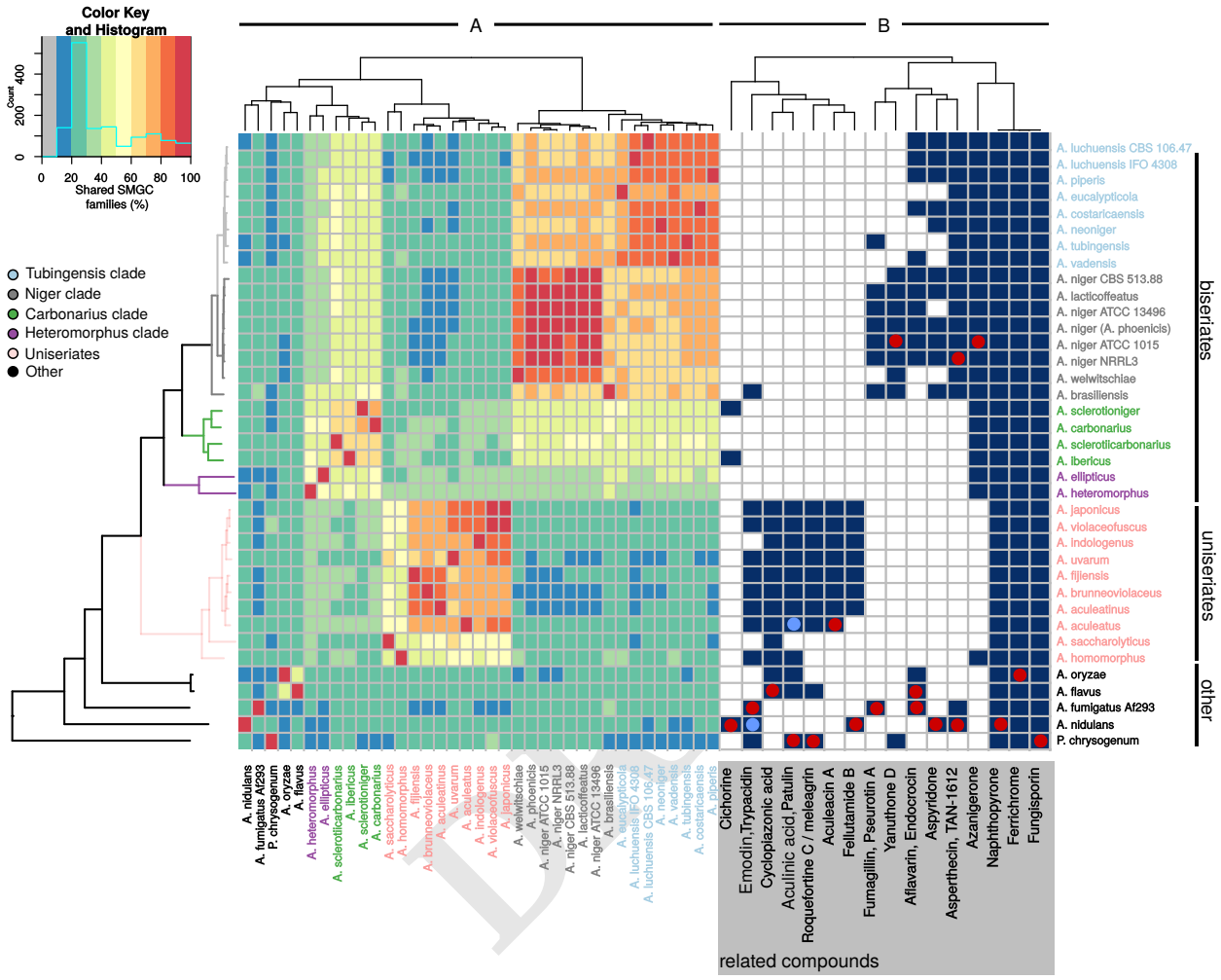


Fig. 1. Heatmap of shared SMGC families and gene clusters linked to compounds. This heatmap contains information on phylogeny of used strains, shared SMGC families and metabolite-linked gene clusters based on MIBiG entries. The row dendrogram represents a whole genome phylogeny of strains in the dataset. (A) Relative amounts of shared SMGC families between species in percent. Here, the presence of SMGC families resulting from our pipeline was compared through all species in the dataset. Percentage is indicated as color gradient in bins of 10% from grey cells (0-10%, not present in dataset) to red cells (90-100%) as shown by the color key. Additionally, a histogram indicates the abundance of different amounts of shared SMGC families, hence, how many comparisons result in low or high similarity respectively. Species self-comparison always results in values of 100%. The key shows that the majority of species to species comparisons indicates a 20-30 % similarity of SMGC families. Within clades, the values are typically above 50%. 80-100% is shared among isolates of the same species and closely related species. The column dendrogram represents a hierarchical clustering of organisms by shared SMGC percent, hence strains clustering together will share a high amount of SMGC. (B) Identification of compound-linked gene clusters based on MIBiG entries. Best hits for MIBiG entries,that suffice the cutoffs, were identified inside families using protein BLAST (red dot). Emodin and aculinic acid were identified manually by sequence identifier (blue dot). Using a guilt by association approach, the whole family of gene clusters is considered to be responsible for the production of a similar metabolite. The heatmap column dendrogram is clustered hierarchically based on presence of compound-linked gene clusters, hence, organisms with shared known SMGC will cluster together.

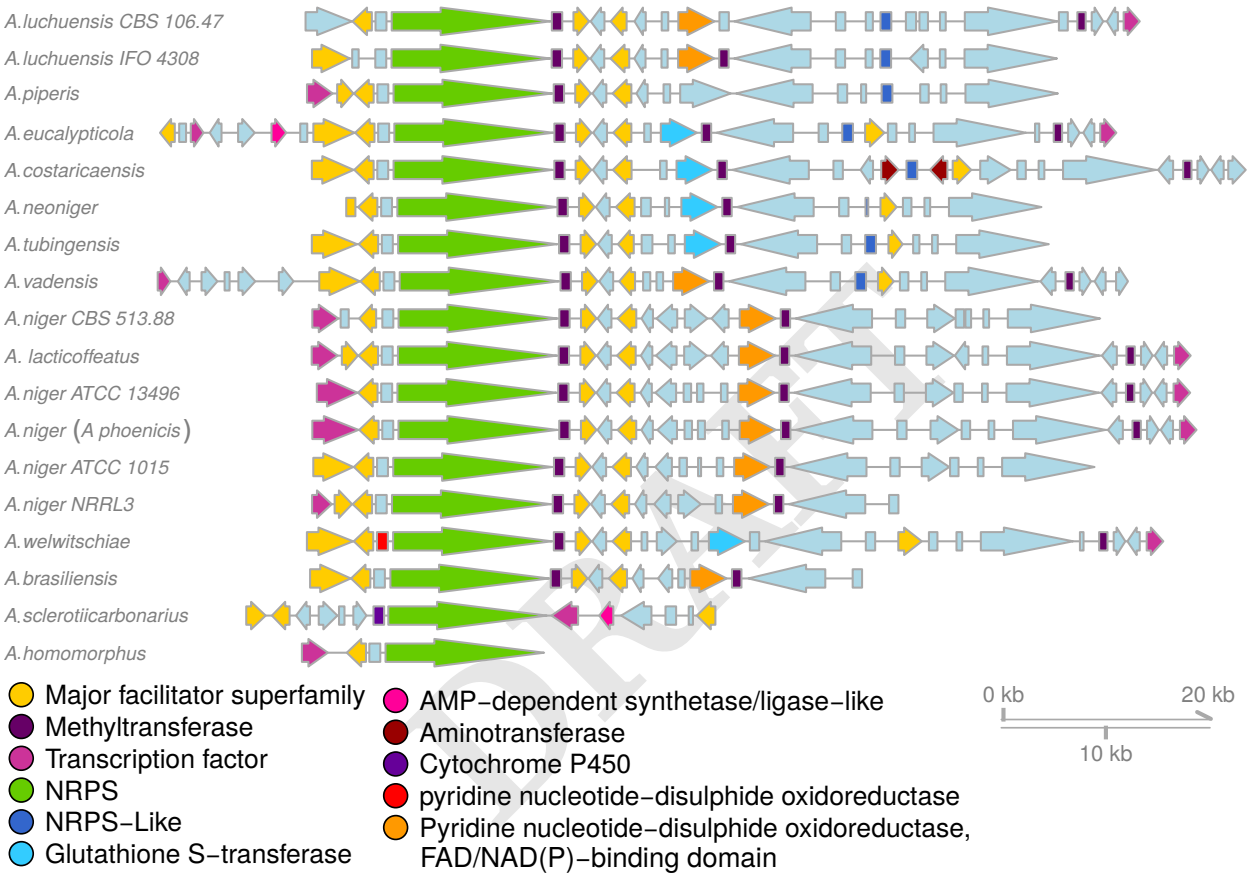


Fig. 2. Predicted SMGC family for malformin producing gene clusters. Intepro annotations are indicated by color. The shown family contains NRPS encoding genes and tailoring enzymes which fulfilled our criteria. We hypothesize that genes coding for glutathione-S-transferases or pyridine-disulphide oxidoreductases are responsible for the disulphide bond in malformin (structure in Fig. 4)

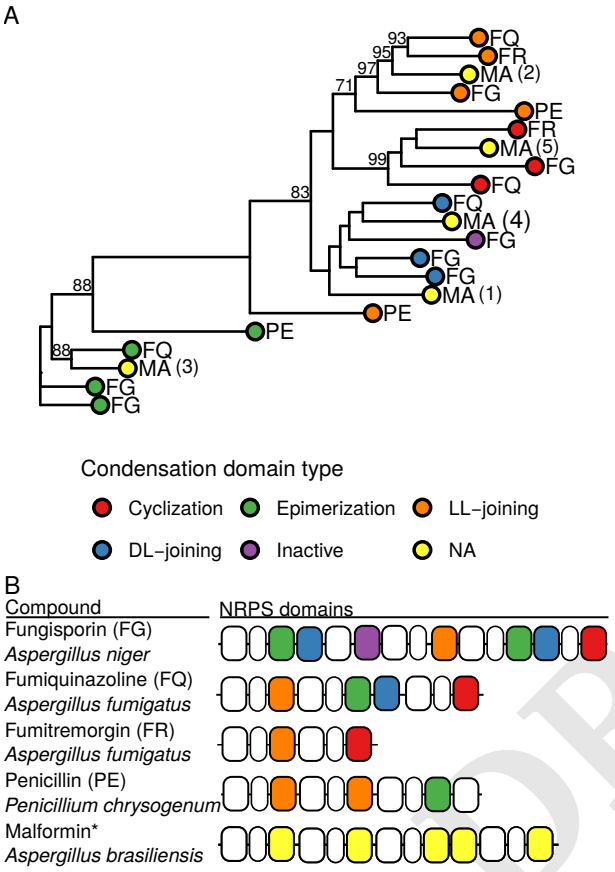


Fig. 3. Classification of condensation domains inside the predicted NRPS responsible for Malformin synthesis. A: Approximate Maximum likelihood phylogeny of condensation domain amino acid sequences. Sequences of condensation domains with known activities from fungisporin (FG), fumiquinazolines (FQ), fumitremorgin (FR) and penicillin (PE) were used to infer activities of condensation domains in the predicted malformin (MA) producing NRPS. The tree was generated from 60% of conserved aligned columns and bootstrapped 1000 times. Bootstrap values over 70 are shown next to their node. The analysis shows distinct clusters corresponding to functions of condensation domains supported by high bootstrap support. B: Schematic for used NRPS proteins. Condensation domains are highlighted according to their function as depicted in the legend (NA: not available). Adenylation and pcp domains are represented by white cells.

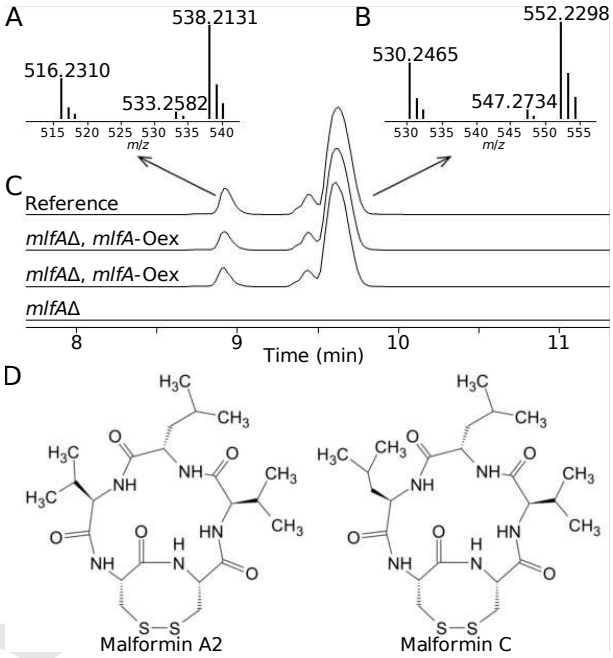


Fig. 4. Extracted Ion Chromatograms (EIC) for malformin overexpressing (*mlfA*Δ, *mlfA*-Oex) and malformin knock-out (*mlfA*Δ) strains. **A** and **B** show MS spectra of detected adducts $[M^+H]^+$, $[M^+NH_4]^+$ and $[M^+Na]^+$ for the peaks displayed in **C** showing merged EICs of the six adducts ($\pm$ 0.005 Da) in the reference strain (*akuA* Δ ::AFL*pyrG*), *mlfA*Δ- *mlfa*-Oex (*mlfA*Δ IS1::P_{gdpA}-*mlfA*) and *mlfA* deletion strain (*mlfA*Δ). **A** reveals the peak at RT 8.9 min contains calc. m/z 516.2310, 533.2582, 538.2131, corresponding to adducts of low-mass malformins, e.g. A2 and C, respectively (**D**). The two peaks at RT 9.4-9.7 min contain the adducts of high-mass malformins, calc. m/z 530.2465, 547.2734, 552.2298 (**B**), where the largest peak at RT 9.7 min represents malformin C as determined by comparison to a reference standard of malformin C (**D**). The small peak at RT 9.4 min denotes another of the high-mass malformin (e.g. malformin A1, B1, B3, B4) (50). In **C**, the vertical axis displaying MS counts is not shown, however the intensity of the tallest peak is approximately 2×10^6.

52. Pel HJ, et al. (2007) Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88. *Nature biotechnology* 25(2):221–31.
53. Khaldi N, Collemare J, Lebrun MH, Wolfe KH (2008) Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. *Genome biology* 9(1):R18.
54. Nielsen ML, et al. (2011) A genome-wide polyketide synthase deletion library uncovers novel genetic links to polyketides and meroterpenoids in Aspergillus nidulans. *FEMS Microbiol Lett* 321(2):157–166.
55. Andersen MR, et al. (2011) Comparative genomics of citric-acid-producing Aspergillus niger ATCC 1015 versus enzyme-producing CBS 513.88. *Genome Research* 21(6):885–897.
56. Nielsen JC, et al. (2017) Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in Penicillium species. 17044(April).
57. Rausch C, Hoof I, Weber T, Wohlleben W, Huson DH (2007) Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC evolutionary biology* 7:78.
58. Bode HB, Bethe B, Höfs R, Zeeck A (2002) Big effects from small changes: possible ways to explore nature's chemical diversity. *Chembiochem* 3(7):619–627.
59. Charlop-Powers Z, et al. (2016) Urban park soil microbiomes are a rich reservoir of natural product biosynthetic diversity. *Proceedings of the National Academy of Sciences* 113(51):201615581.
60. Ziemert N, et al. (2014) Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. *Proceedings of the National Academy of Sciences of the United States of America* 111(12):1130–9.
61. Csardi G, Nepusz T (2006) The igraph software package for complex network research. *InterJournal* Complex Sy:1695.
62. Pons P, Latapy M (2005) Computing communities in large networks using random walks. *Physics and Society* p. arXiv:physics/0512106.
63. R Core Team (2017) R: A Language and Environment for Statistical Computing.
64. Warnes GR, et al. (2016) gplots: Various R Programming Tools for Plotting Data.
65. Yu G, Smith D, Zhu H, Guan Y, Lam TTY (2017) ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*.
66. Hahne F, Ivanek R (2016) Visualizing Genomic Data Using Gviz and Bioconductor in *Statistical Genomics: Methods and Protocols*, eds. Mathé E, Davis S. (Springer New York, New York, NY), pp. 335–351.
67. Katoh K, Misawa K, Kuma Ki, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* 30(14):3059–3066.
68. Castresana J (2000) Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution* 17(4):540–552.
69. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
70. Jones P, et al. (2014) InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30(9):1236–1240.
71. Sievers F, et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology* 7(1):539.
72. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972–1973.
73. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32(1):268–274.
74. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS (2017) ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates. *Nature Methods* 14(6):587–591.
75. Minh BQ, Nguyen MAT, Von Haeseler A (2013) Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and Evolution* 30(5):1188–1195.
76. Varga J, et al. (2007) Aspergillus brasiliensis sp. nov., a biseriate black Aspergillus species with world-wide distribution. *International Journal of Systematic and Evolutionary Microbiology* 57:1925–1932.
77. Frisvad JC, Samson RA (2004) Polyphasic taxonomy of Penicillium subgenus Penicillium: A guide to identification of food and air-borne terverticillate Penicillia and their mycotoxins. *Studies in Mycology* 2004(49):1–173.
78. Kildgaard S, et al. (2014) Accurate dereplication of bioactive secondary metabolites from marine-derived fungi by UHPLC-DAD-QTOFMS and a MS/HRMS library. *Marine Drugs* 12(6):3681–3705.

Chapter 4

Genus level analysis of PKS-NRPS and NRPS-PKS hybrids reveals their origin in Aspergilli

The previous chapters were dealing with secondary metabolite gene clusters as a whole and focusing on creating families that produce similar compounds. In this chapter one enzyme class is being investigated more closely to find dynamics on the domain level between SM genes. Phylogenetic studies of secondary metabolite enzymes like polyketide synthases (PKS) and non-ribosomal peptide synthetases (NRPS) have elucidated their evolutionary dynamics and origins. Hybrids of these enzymes, PKS-NRPS and NRPS-PKS enzymes, however, have not been the focus of these studies — neglecting their evolution. In this study, we generated a phylogeny focusing on hybrids to investigate their diversity and dynamics in the genus *Aspergillus*. Furthermore, we relate hybrids to PKSs and NRPSs using their adenylation and ketoacylsynthase domains, since we hypothesize that hybrids can, as opposed to previous studies Gallo et al. (2013), be polyphyletic. Based on evidence that fungi have acquired hybrids from bacteria and fungi by lateral transfer(Lawrence et al., 2011; Khaldi et al., 2008), we blasted fungal hybrids against the NCBI non-redundant protein database to find further evidence for horizontal gene transfer. Our analysis indicates that hybrids are frequently transferred between fungi and one class of hybrids originated in bacteria. Our results highlight the dynamics of this enzyme class and will serve as guideline for recombination experiments to form new chimeras.

RESEARCH

Genus level analysis of PKS-NRPS and NRPS-PKS hybrids reveals their origin in Aspergilli

Sebastian Theobald*, Tammi C. Vesth and Mikael R. Andersen

Abstract

Background:

Filamentous fungi produce a vast amount of bioactive secondary metabolites (SMs) synthesized by e.g. hybrid polyketide synthase-nonribosomal peptide synthetase enzymes (PKS-NRPS; NRPS-PKS). While their domain structure suggests a common ancestor with other SM proteins, their evolutionary origin and dynamics in fungi are still unclear. Recent rational engineering approaches highlighted the possibility to reassemble hybrids into chimeras — suggesting molecular recombination as diversifying mechanism.

Results:

Phylogenetic analysis of hybrids in 39 strains lets us describe their dynamics throughout the genus *Aspergillus*. The tree topology indicates that three groups of PKS-NRPS hybrids as well as one group of NRPS-PKS developed independently from each other in Aspergilli. Comparison to other SM genes lead to the conclusion that hybrids in *Aspergilli* have several PKS ancestors; in contrast, hybrids are monophyletic when compared with NRPSs — with the exception of a small group of NRPSs. Our analysis also revealed that NRPS-likes are derived from NRPSs at many different events. An extended phylogenetic analysis including multiple bacterial and fungal taxa revealed multiple ancestors of hybrids. Homologous hybrids are present in all sections which suggests frequent horizontal gene transfer between genera and a finite number of hybrids in fungi.

Conclusion:

Phylogenetic distances between hybrids provide us with evidence for their evolution: Large inter-group distances indicate multiple independent events leading to hybrid development, while short intra-group distances of hybrids from different taxonomic sections indicate frequent horizontal gene transfer. Our results are further supported by adding bacterial and fungal genera. Presence of related hybrid genes in all *Ascomycetes* suggests a frequent horizontal gene transfer between genera and a finite diversity of hybrids — also explaining their scarcity. The provided insights into relations of hybrids and other SM genes will serve in rational design of new hybrid enzymes.

Keywords: Aspergillus; PKS-NRPS hybrids; Secondary metabolites; Gene clusters

Background

Secondary metabolites (SMs), non-growth associated compounds, have been subject to research efforts due to their wide range of bioactivities. Polyketides like sterigmatocystin and aflatoxin, two potent mycotoxins, cause food spoilage; while others like the cholesterol lowering lovastatins can be used as medical drugs. Many SMs are promising leads for anti-cancer drugs as e.g. the non-ribosomal peptides malformins [1]. The enzymes classes producing these distinct compounds — polyketide synthases (PKSs) and non-ribosomal peptide synthetases (NRPSs) — can also be fused into a PKS-NRPS or NRPS-PKS hybrid, creating a chimeric compound. The products of hybrids, e.g. cyclopiazonic acid, pyranonigrin and cytochalasin show a wide range of bioactivities [2–4].

Evolutionary events leading to new enzymes and hence compounds have been described in detail for PKSs and NRPSs. Exchange of initiation modules for

*Correspondence: setd@bio.dtu.dk
Department of Biotechnology and Biomedicine, Technical University of Denmark, Anker Engelunds vej 1, Kgs. Lyngby, DK
Full list of author information is available at the end of the article

modification of primer units, module duplication, horizontal gene transfer and acquisition of tailoring enzymes diversified polyketides [5–7]. Other studies suggest a burst of PKS duplications in the early *Pezizomycotina*, a predeccor of mainly *Ascomycota*, [8] as major driver for PKS diversity. Similarly, phylogenomic studies of NRPSs suggest a horizontal gene transfer from bacteria to fungi [9]. suggest mono and bimodular NRPSs to be most abundant in bacteria, while euascomycetes show multimodular NRPSs in response to gain and loss of domains.

In contrast, hybrids have been neglected by phylogenetic studies, although their combination of NRPS and PKS domains suggests an interesting evolutionary history. The existing studies only focused on known hybrids [10]. Lawrence et al [11] have shown that a single *Cochliobolus heterostrophus* NRPS-PKS hybrid gene originates from Burkholderiales; which they suggest to have taken place in the early development of the Pezizomycotina.

A model for the avirulence factor *ACE1* gene [12] by Khaldi et al. [13] shows gene duplication, loss, and horizontal gene transfer — a common event between fungi [14] — as driver in diversification of *ACE1* hybrids. With an extended dataset, we sought to identify the origins of hybrids in *Aspergillus*, providing insight into molecular evolution — the basis of chemical innovation. Understanding hybrid evolution and diversity will put rational engineering of these proteins into our grasp and pave the way for new bioactive compounds.

Thus, we sought to investigate the phylogenetic dynamics of PKS-NRPS and NRPS-PKS in the SM rich genus *Aspergillus* and relate them to other classes of SM genes. Furthermore, we identify origins of hybrids in bacteria and other fungal genera.

Results

Genus wide analysis identifies independent groups of hybrids

Recent work has highlighted the dynamics of SM genes in fungi. NRPS are subject to duplication and less, frequent domain gain and loss, with fungi specific groups emerging [15, 16]; and PKSs show great diversity in fungi as well [8]. Hence we were interested why PKS-NRPS and NRPS-PKS hybrids are relatively few in *Aspergillus* species compared to PKSs and NRPSs and whether we find the evolutionary events which led to hybrid evolution. Due to apparent independent parts, we expect a fusion of NRPS and PKS occured during early fungal evolution.

In order to investigate this, we created a maximum likelihood phylogeny of hybrid proteins from Aspergilli of section *Nigri* (including biseriates and uniseriates), *Circumdati*, *Candidi*, *Flavi*, *Fumigati*, *Ochracerosii*, *Terrei* and *P. chrysogenum* to cover a wide range of *Eurotiomycetes* (Fig. 1. In the phylogeny, NRPS-PKS and PKS-NRPS hybrids form several distinct groups in the phylogeny, with the PKS-NRPS orientation being more abundant in Aspergilli (Fig. 1). One group including the cytochalasin-associated hybrid gene *ccsA* only include few species and show large phylogenetic distance to other hybrids, indicating that they are rare in Aspergilli. The second group includes hybrids which are conserved in biseriate Nigri species, but also *A. homomorphus*, *A. clavatus* *A. campestris* and *A. ochraceroseus*. The tree topology indicates this as common hybrid duplication in Aspergilli through its conservation in many species. The short phylogenetic distance of *A. ibericus* and *A. sclerotiicarbonarius* hybrids to *A. campestris* and *A. ochraceroseus* hybrids is remarkable, since these species are from different sections. Additionally, it shows how hybrids conserved in one subgroup, the biseriates of section *Nigri*, can be found in *A. homomorphus* — a member of uniseriates. Furthermore, reaquisition of related hybrids appears to be common since *A. ibericus* and *A. steynii* contain duplications with larger phylogenetic distances (Aspibe1_443386 and Aspibe1_469268; Aspste1_454498 and Aspste1_477231).

The third group contains pseurotin A and isoflavipucine linked hybrids. The presence of these hybrids in two sister clades is surprising since they use different amino acids and produce distinct compounds. However, the broad substrate acceptance of isoflavipucine, which has been shown to accept many different substrates to create 63 diverse compounds [17], supports a common origin. A fourth group contains hybrids related to chaetoglobosin, aspyridone and cyclopiazonic acid hybrids. It is surprising that chaetoglobosin and cytochalasin hybrids show exceptional distance in our analysis. This suggests that, although the hybrids are similar in their biosynthetic activity, they evolved individually.

Following are four groups which are unrelated to the other hybrids, two consisting of PKS-NRPS, including pyranonigrin related hybrids, and two groups of NRPS-PKS orientation. Pyranonigrin related hybrids are, with the exception of one hybrid from *A. steynii*, unique for section *Nigri* (at least in the scope of our dataset). Notably, NRPS-PKS hybrids are rare among the analyzed species and are only present in a few species: the biseriates of section *Nigri*, *A. indologenus*, *A. steynii*, *A. campestris*, and *Penicillium chrysogenum*. Their absence in other species and low number points towards recent acquisition by horizontal gene transfer (HGT) of all NRPS-PKS. The phylogeny indicates two major groups of NRPS-PKS hybrids and two hybrids as outgroups (Aspcam1_323099

and Aspste1_418130). While one major group is biseriate specific, the other group consists of *P. chrysogenum* (Pench1_85311), hybrids from biseriates, and *A. indologenus*. *A. luchuensis* and *A. piperis* are the only species that carry hybrids from both major groups, pointing towards a HGT before their speciation, or retention of a hybrid. The position of *P. chrysogenum* hybrid in the phylogeny points towards HGT as well. The NRPS-PKS oriented hybrids show very large distance to the rest of hybrids in the phylogenetic tree, caused by the different orientation and bias of the alignment towards the more prominent PKS-NRPS oriented hybrids. Hence, we created alignments of single adenylation and ketosynthase domains which would provide an unbiased view of the relation of NRPS-PKS and PKS-NRPS hybrids.

Genus wide analysis provides evidence for HGT

Horizontal gene transfer of hybrids has been observed for fungi of different genera. With the ML phylogeny established, it was an obvious step to extend our analysis for detection of potential horizontal gene transfer. If all hybrids were inherited vertically, we would expect for a given synthetase that the best homologs (nearest neighbour in the tree) originate from the same section as the synthetase itself. To find the best homologs of hybrids for each species, we extracted distances of hybrids from the ML phylogeny and classified them according to origin (Fig. 1). This analysis works best with the hybrid rich section *Nigri*, since we cover most species here, but also the other *Aspergillus* species and *P. chrysogenum* provide insights into hybrid dynamics.

Our analysis reveals that biseriates of section *Nigri* contain mostly conserved groups of hybrids, but can contain some hybrids derived from other parts of the phylogeny (Fig. 1). *A. sclerotiicarbonarius* contains a related hybrid to the aspyridone gene from *A. nidulans* and *A. heteromorphus* contains a majority of uniseriate hybrid homologs. *A. ibericus* contains one third of hybrid homologs from other sections. *A. sclerotioniger* contains a uniseriate homolog (protein id 605326), located in a group together with hybrids from *A. heteromorphus*, a biseriate, *P. chrysogenum* and *A. clavatus* ccsA. Thus, we predict this hybrid to produce sclerotionigrin [18], a related compound to cytochalasins produced by ccsA. It is suprising however that *A. sclerotioniger* 605326 shows a shorter phylogenetic distance to the *A. heteromorphus* hybrid. Hybrids have been shown previously to be subject to loss and reaquisition. This conservation could be the results of a reaquisition of a hybrid from uniseriates to biseriates — the apparent absence of ccsA homologs in section *Nigri* supports a deletion event.

Uniseriates of section *Nigri*, with species containing less hybrids than the biseriates, show foreign homologs as well. *A. saccharolyticus*, a uniseriate, contains one hybrid from biseriate species and one homolog which shows high conservation to a hybrid from *A. steynii* and *A. oryzae* (Fig. 1). The latter is responsible for cyclopiazonic acid (CPA) synthesis, a mycotoxin [19].

In the dataset, we included only few non *Nigri* species with often only one representative per section. Hence, figure 1 will show that all hybrids in these species have homologs in species of other sections than their own. This is of course, not the case, the analysis however still gives a good indication of the origin of their hybrids. As in the case for the isoflavipucine hybrid from *A. terreus* and a hybrid from *A. campestris*. Here, a short phylogenetic distance indicates HGT between these species rather than hybrid conservation. Although not derived by HGT, but still worth mentioning are hybrids from *A. steynii*. Its hybrids are related to almost every subgroup of hybrids in the dataset.

A. steynii and section *Nigri* species contain a large number of diverse hybrids which are related to most subgroups in the dataset. If new lineages of hybrids would frequently emerge throughout sections we would expect more section specific hybrids and *A. steynii* as well as *Nigri* species would cover less of the hybrid groups. This suggests that the evolutionary events leading to hybrid generation happened before species diversification in the genus *Aspergillus*. Another observation is that closely related hybrids are present in many phylogenetically distant sections. This suspects that diversification of hybrids is happening through recombination events after HGT of hybrids. Additionally, we expect that NRPS-PKS hybrids were either derived by joining of independent NRPS and PKS genes or acquired independently from another source, since they show large phylogenetic distance to PKS-NRPS hybrids. Since the phylogenetic distance could be biased by the amount of structurally similar PKS-NRPS hybrids we created further comparisons on basis of single domains. There are however, intrinsically large distances between PKS-NRPS distances as well which we wanted to investigate further.

PKS analysis shows common ancestors for PKSs and hybrids

Since intrinsically, hybrids did show large phylogenetic distances (Fig. 1), we hypothesized their origin from related SM genes. Previous studies prove hybrid parts as exchangeable, hence, we proposed that other SM genes could join together in filamentous fungi to form a hybrids. In order to study this, we created a ML phylogeny of 1369 ketosynthase (KS) domains of PKS-like, PKS and hybrids to elucidate their phylogenetic relations (Fig. 2).

The tree topology shows multiple groups consisting of PKS only, mixed PKS and hybrids and PKS-likes. PKS-likes form two unrelated groups, suggesting that they are largely unrelated to PKSs. Hybrids are separated into two groups. NRPS-PKS hybrids are located as sister clade to 6-methylsalicylic acid (6-MSA) PKS related genes — including PKSs for synthesis of yanuthones, terreic acid and patulin (Fig. 3 right panel). PKS-NRPSs are clustering together with other PKSs that frequently break into hybrid clades and separate known examples from each other (Fig. 3 left panel). Thus, we suggest these PKSs and PKS-NRPS hybrids to have common ancestors in fungi. PKS linked to citreoviridin and pyripyroene as sister clade to hybrids (the pyriporene hybrid has and adjacent adenylation (A) domain in its cluster). Thus these PKSs could be the ideal precursor for the molecular evolution of hybrids.

In summary, hybrids do not form a monophyletic clade inside the ML phylogeny. Hence we can hypothesize that PKSs and hybrids had common ancestors — distinct ones for NRPS-PKS and PKS-NRPS genes. Additionally, the analysis shows that NRPS-PKS and PKS-NRPS hybrids are unrelated as indicated by the phylogeny of hybrids (Fig. 1).

Phylogeny of NRPSs and hybrids reveals monophyletic clade of hybrids

Hybrids incorporate amino acids (e.g. tyrosine in case of the cytotoxic aspyridone or L-phenylalanine in case of cytochalasins) into compounds in a manner similar to NRPS and NRPS-likes. Thus we sought to investigate the phylogenetic relationship of these proteins.

We created a maximum likelihood tree of 2428 adenylation domains from NRPS, NRPS-like and hybrid proteins which led to mostly monophyletic groups (Fig. 4). NRPS-likes form two groups which are monophyletic. Other groups comprise NRPS which appear to have a common ancestor, with few NRPS-likes forming a sister clade. Hybrids form a monophyletic group, they are however located in a sister clade with a group of NRPS and NRPS-likes conserved in uniseriate *Nigri* species (Fig. 5, NRPS homologs are also found in *A. heteromorphus* and *A. ellipticus*). These proteins could possibly have a common ancestor.

Extended analysis of hybrids shows two events leading to hybrid development

Since the comparison of A domains between NRPS, NRPS-likes and hybrids did not deliver the same amount of ancestors we observed for PKS, we sought to extend our search space to the National Center for Biotechnology Information (NCBI) non-redundant protein database. We used protein blast on the NCBI nr database to find homologs of *Aspergillus* hybrid genes. Adenylation domains from 288 best hits were extracted and added to the *Aspergillus* hybrid adenylation domain set. Subsequent alignment and ML analysis generated the phylogeny in Fig. 6.

Best blast hits were mostly originating from *Ascomycete* classes: *Dothideomycetes*, *Eurotiomycetes*, *Leotiomycetes*, *Sordariomycetes*, *Xylonomycetes*, one *Orbilliomycete*, and one *Exobasidiomycete* hybrid were included. Hybrids from bacterial classes include *Proteobacteria*, *Terrabacteria* and *Planctomycetes*. We find fungal sequences distributed throughout the tree, and although many ascomycete taxa are included, the tree topology indicates that hybrids are conserved throughout these taxa. Certainly, our blast search might bias the tree topology, nonetheless, if PKS and NRPS would recombine frequently in fungi, we would expect more intermediates. There are some NRPS, mostly from *Sordariamycetes* and *Dothideomycetes*, which are related to PKS-NRPS hybrids. These could be remnants of ancestral NRPS which have been donors for hybrids in fungi. The majority of the tree consists of PKS-NRPS hybrids, while NRPS-PKS hybrids from fungi and bacterial NRPSs and hybrids (from *Terrabacteria* and *Proteobacteria*), are coclustering in one location (Fig. 6). Inside the cluster, we can identify the thanamycin hybrid gene from *Pseudomonas sp.* SHC52, a lipopeptide. What's more, we can identify hybrids KPC78190.1, APD71785.1, WP_023586037.1 from *Streptomyces sp.* in a sister clade to hybrids from multiple fungal genera. This indicates that lipopeptides could be a progenitor of NRPS-PKS in general and a hybrid from *Streptomyces* could be horizontally transferred to fungi — giving rise to NRPS-PKS hybrids in fungi.

The phylogeny also shows related hybrids of different sections in the same branch, as in the case for genes of aspyridone and fumosorinone — two compounds similar in structure [20]. This emphasizes that the structural diversity of hybrids throughout *Ascomycetes* might be limited. PKS-NRPS homologs of *Aspergilli* are co-clustering with many hybrids from *Sordariamycetes* and *Eurotiomycetes*. Thus, the distances of pyranonigrin associated hybrids observed earlier (Fig. 1) can now be explained with the added dataset. The *A. ellipticus* hybrid (460246) is clustering closer together with Sordariamycetes; the same for *A. steynii* which carries an Eurotiomycete related hybrid. Recurrence of the same genera emphasizes that hybrids might be limited in fungi, which is why there are usually so few.

Discussion

Analysis of the whole SM protein repertoire of a genus wide dataset led us to discover dynamics between

NRPSs, NRPS-likes, PKSs, PKS-likes and NRPS-PKS as well as PKS-NRPS hybrids. While previous studies included hybrids related to known examples [10], focused on NRPS domains [9], or where considering single hybrids for analysis [11, 13] we combined analysis of A and KS-domains through all hybrids in a genus wide species set and related them to other SM enzymes to focus on their evolutionary history.

Gallo et. al [10] also showed that the hybrid and NRPS adenylation domains are monophyletic — which our analysis supports. Yet, we could identify a sister clade of NRPS and NRPS-likes which groups together with hybrids. These genes are valuable leads in the investigation of SM gene evolution in the genus *Aspergillus*. Additionally, one common ancestor of PKS and Hybrid PKS-NRPS had been hypothesized [10]. Our results indicate multiple PKSs which could be the ancestor of hybrids. We can confirm that despite structural (or biosynthetic) similarity of cytochalasins and chaetoglobosisn, the genes for their synthetases have a distinct phylogenetic history — ccsA, the cytochalasin hybrid, is more closely related to the equisetin hybrid than to the chaetoglobosin hybrid.

Previous studies indicated that NRPS-PKS are of bacterial origin, while PKS-NRPS due to their abundance in fungi are of fungal origin [11]. Our analysis shows that only A and KS domains of NRPS-PKS hybrids have similar phylogenetic histories in fungi (as indicated in [11]). The extended comparison to hits from the NCBI non redundant protein database in this study revealed that the collective of NRPS-PKS in *Ascomycetes* is related to bacterial hybrids and lipopeptide synthetases, suggesting bacterial origin. PKS-NRPS hybrids on the other hand, show similarity to different fungal PKSs in the ML phylogeny. Hence, we suspect that A and KS domains have a different phylogenetic history. Their monophyly in A domain comparisons suggests that one NRPS ancestor was able to recombine with different PKSs. Hence our analysis shows that multiple events rather than one event gave rise to hybrid evolution.

According to previous studies, the *ccsA* ancestor was duplicated in *Ascomycete* development, lost in Aspergilli, and reacquired by *A. clavatus* from *Magnaporthe grisea* [13]. In our analysis of SM genes throughout 39 *Aspergillus* strains — with addition of best hits from the NCBI non redundant protein database — we can further provide evidence for HGT of hybrids through ascomycetes. Biseriate species *A. sclerotioniger*, *A. heteromorphus* as well as uniseriate species *A. saccharolyticus*, all of which belong to section *Nigri*, contain conserved *ccsA* homologs (Fig. 1). *A. sclerotioniger* is producer of sclerotionigrin, a SM related to cytochalasins from *A. clavatus*. Duplication, loss and reacquisition of related hybrids appears to be common since *A. ibericus* and *A. steynii* contain duplications with larger phylogenetic distances (Aspibe1_443386 and Aspibe1_469268; Aspste1_454498 and Aspste1_477231). This duplication, loss, and reacquisition pattern is similar to the one proposed for *ace1*. *Ace1* is duplicated during *Sordariamycete* evolution to *ace1* and *syn2*. In an HGT event, *syn2* is transferred from *Magnaporthe grisea* to *A. clavatus* resulting in *ccsA* — an organism which lost the original *ace1* ancestor [13]. Additionally, we find several cases where phylogenetic distances to hybrids of other sections are shorter than to hybrids of their own section. For example, *A. heteromorphus*, although biseriate, contains mostly hybrid homologs from from uniseriate species and *A. saccharolyticus* contains a homolog to a hybrid from of *A. flavus*. This is further supported by our extended analysis using additional hybrids from other fungi and bacteria. *A. ellipticus* and *A. steynii* show homologs to *Sordariomycetes* and *Eurotiomycetes*, respectively — indicating a loss and reaquisition of a pyranonigrin related hybrid. The phylogenetic relationship of hybrids points to few events which yielded novel hybrids, the phylogenetic distances however, exclude a loss only diversification. We suggest that hybrids are frequently transferred, hence the short distances betweeen hybrids of distantly related species and the finite number of clades/ their low amount.

Aspergilli show vast amounts of SM genes; hybrids however, are the minority (Table 1). These low counts, their dynamics and their unrelated groups 1 pose the question whether independent events lead to their evolution and whether their diversity is finite.

Hybrids were subject to many engineering efforts since their structure suggested inpendendent units, which could be recombined. TenS and DmbS, part of the hybrids producing the similar compounds tenellin and desmethylbassianin, respectively, could be fused together and resulted in the production of tenellin A [21]. Nielsen et. al [22] created a fusion of CcsA and Syn2 resulting in niduporthin, a novel chimeric compound. Despite these successful recombinations, creating a chimera of hybrids remained challenging as other studies showed a limitation to recombination efforts [23]. Our study will facilitate further efforts to engineer hybrids since we could show their dynamics inside the genus *Aspergillus* and relate it to other fungal genera. Especially NRPS and PKS which have been identified in this study as hybrid related might be amenable to modification.

Methods

Data collection

Protein sequences and smurf annotations for *Aspergillus* and *Penicillium* species were downloaded from JGI (http://genome.jgi.doe.gov/).

Genetic dereplication

Secondary metabolite proteins were annotated with known compounds through BLAST against examples from the Minimum information on biosynthetic gene clusters (MIBiG) database [24].

Identifying best hits in the NCBI non redundant protein database

Adenylation domains of hybrids were blasted against the NCBI non redundant protein database (downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/db/) using protein basic local alignment search tool (BLAST) [25]. The best 15 hits which suffice a query coverage cutoff of 80% were retained. Taxonomic labels for best hits were added using ETE 3 toolkit [26].

Protein domains, alignment and maximum liklihood analysis

InterproScan5 (version 5.22-61.0) [27] was used to identify domains on protein sequences. Domains of the same type which were under 350 amino acids long and less than 100 amino acids apart were merged before proceeding. Sequences were handled using Biopython [28]. Resulting domain sequences were aligned using Clustal Omega [29] and cut using trimal [30]. In case of the full hybrid protein tree, full protein sequences were aligned and trimmed. IQ-tree [31] was used on aligned sequences using Model Finder Plus [32] and 1000 times ultra fast bootstrap [33].

Finding best homologs in trees

A distance matrix was extracted from the ML tree using the cophenetic function of the ape package [34] in R [35]. Best homologs were plotted using ggplot2 [36]

Visualization

Phylogenetic trees were visualized using ggtree [37]. ggstance [38]. matplotlib [39]. Gene clusters were plotted using Easyfig [40].

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

ST, TCV, and MRA acknowledge funding from The Villum Foundation, grant VKR023437

References

1. Wang, J., Jiang, Z., Lam, W., Gullen, E.A., Yu, Z., Wei, Y., Wang, L., Zeiss, C., Beck, A., Cheng, E.C., Wu, C., Cheng, Y.C., Zhang, Y.: Study of malformin C, a fungal source cyclic pentapeptide, as an anti-cancer drug. PLoS ONE **10**(11), 1–19 (2015). doi:10.1371/journal.pone.0140069
2. Awakawa, T., Yang, X.L., Wakimoto, T., Abe, I.: Pyranonigrin E: A PKS-NRPS hybrid metabolite from aspergillus niger identified by genome mining. ChemBioChem **14**(16), 2095–2099 (2013). doi:10.1002/cbic.201300430
3. Tokuoka, M., Seshime, Y., Fujii, I., Kitamoto, K., Takahashi, T., Koyama, Y.: Identification of a novel polyketide synthase-nonribosomal peptide synthetase (PKS-NRPS) gene required for the biosynthesis of cyclopiazonic acid in Aspergillus oryzae. Fungal Genetics and Biology **45**(12), 1608–1615 (2008). doi:10.1016/j.fgb.2008.09.006
4. Casella, J.F., Flanagan, M.D., Lin, S.: Cytochalasin D inhibits actin polymerization and induces depolymerization of actin filaments formed during platelet shape change. Nature **293**(5830), 302–305 (1981). doi:10.1038/293302a0
5. McDaniel, R., Ebert-Khosla, S., Hopwood, D.a., Khosla, C.: Rational design of aromatic polyketide natural products by recombinant assembly of enzymatic subunits. (1995). doi:10.1038/375549a0
6. Ridley, C.P., Lee, H.Y., Khosla, C.: Evolution of polyketide synthases in bacteria. Proceedings of the National Academy of Sciences of the United States of America **105**(12), 4595–600 (2008). doi:10.1073/pnas.0710107105
7. Throckmorton, K., Wiemann, P., Keller, N.P.: Evolution of Chemical Diversity in a Group of Non-reduced Polyketide Gene Clusters: Using Phylogenetics to Inform the Search for Novel Fungal Natural Products vol. 7, pp. 3572–3607 (2015). doi:10.3390/toxins7093572
8. Koczyk, G., Dawidziuk, A., Popiel, D.: The distant siblings - A phylogenomic roadmap illuminates the origins of extant diversity in fungal aromatic polyketide biosynthesis. Genome Biology and Evolution **7**(11), 3132–3154 (2015). doi:10.1093/gbe/evv204
9. Bushley, K.E., Turgeon, B.G.: Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. BMC evolutionary biology **10**(1), 26 (2010). doi:10.1186/1471-2148-10-26
10. Gallo, A., Ferrara, M., Perrone, G.: Phylogenetic study of polyketide synthases and nonribosomal peptide synthetases involved in the biosynthesis of mycotoxins. Toxins **5**(4), 717–742 (2013). doi:10.3390/toxins5040717
11. Lawrence, D.P., Kroken, S., Pryor, B.M., Arnold, A.E.: Interkingdom gene transfer of a hybrid NPS/PKS from bacteria to filamentous Ascomycota. PloS one **6**(11), 28231 (2011). doi:10.1371/journal.pone.0028231
12. Böhnert, H.U., Fudal, I., Dioh, W., Tharreau, D., Notteghem, J.-L., Lebrun, M.-H.: A putative polyketide synthase/peptide synthetase from Magnaporthe grisea signals pathogen attack to resistant rice. The Plant cell **16**(9), 2499–513 (2004). doi:10.1105/tpc.104.022715
13. Khaldi, N., Collemare, J., Lebrun, M.-H., Wolfe, K.H.: Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. Genome biology **9**(1), 18 (2008). doi:10.1186/gb-2008-9-1-r18
14. Fitzpatrick, D.a.: Horizontal gene transfer in fungi. FEMS Microbiology Letters **329**(1), 1–8 (2012). doi:10.1111/j.1574-6968.2011.02465.x
15. Bushley, K.E., Ripoll, D.R., Turgeon, B.G.: Module evolution and substrate specificity of fungal nonribosomal peptide synthetases involved in siderophore biosynthesis. BMC evolutionary biology **8**(1), 328 (2008). doi:10.1186/1471-2148-8-328
16. Bushley, K.E., Turgeon, B.G.: Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. BMC evolutionary biology **10**, 26 (2010). doi:10.1186/1471-2148-10-26
17. Qiao, K., Zhou, H., Xu, W., Zhang, W., Garg, N., Tang, Y.: A fungal nonribosomal peptide synthetase module that can synthesize thiopyrazines. Organic Letters **13**(7), 1758–1761 (2011). doi:10.1021/ol200288w
18. Petersen, L.M., Bladt, T.T., Dürr, C., Seiffert, M., Frisvad, J.C., Gotfredsen, C.H., Larsen, T.O.: Isolation, structural analyses and biological activity assays against chronic lymphocytic leukemia of two novel cytochalasins - Sclerotionigrin A and B. Molecules **19**(7),

9786–9797 (2014). doi:10.3390/molecules19079786
19. Sorenson, W.G., Tucker, J.D., Simpson, J.P.: Mutagenicity of the tetramic mycotoxin cyclopiazonic acid. Applied and Environmental Microbiology **47**(6), 1355–1357 (1984)
20. Liu, L., Zhang, J., Chen, C., Teng, J., Wang, C., Luo, D.: Structure and biosynthesis of fumosorinone, a new protein tyrosine phosphatase 1B inhibitor firstly isolated from the entomogenous fungus Isaria fumosorosea. Fungal Genetics and Biology **81**, 191–200 (2015). doi:10.1016/j.fgb.2015.03.009
21. Heneghan, M.N., Yakasai, A.A., Williams, K., Kadir, K.A., Wasil, Z., Bakeer, W., Fisch, K.M., Bailey, A.M., Simpson, T.J., Cox, R.J., Lazarus, C.M.: The programming role of trans-acting enoyl reductases during the biosynthesis of highly reduced fungal polyketides. Chemical Science **2**(5), 972 (2011). doi:10.1039/c1sc00023c
22. Nielsen, M.L., Isbrandt, T., Petersen, L.M., Mortensen, U.H., Andersen, M.R., Hoof, J.B., Larsen, T.O.: Linker flexibility facilitates module exchange in fungal hybrid PKS-NRPS engineering. PLoS ONE **11**(8), 1–18 (2016). doi:10.1371/journal.pone.0161199
23. Boettger, D., Bergmann, H., Kuehn, B., Shelest, E., Hertweck, C.: Evolutionary Imprint of Catalytic Domains in Fungal PKS-NRPS Hybrids. ChemBioChem **13**(16), 2363–2373 (2012). doi:10.1002/cbic.201200449
24. Medema, M.H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J.B., Blin, K., De Bruijn, I., Chooi, Y.H., Claesen, J., Coates, R.C., Cruz-Morales, P., Duddela, S., Düsterhus, S., Edwards, D.J., Fewer, D.P., Garg, N., Geiger, C., Gomez-Escribano, J.P., Greule, A., Hadjithomas, M., Haines, A.S., Helfrich, E.J.N., Hillwig, M.L., Ishida, K., Jones, A.C., Jones, C.S., Jungmann, K., Kegler, C., Kim, H.U., Kötter, P., Krug, D., Masschelein, J., Melnik, A.V., Mantovani, S.M., Monroe, E.A., Moore, M., Moss, N., Nützmann, H.W., Pan, G., Pati, A., Petras, D., Reen, F.J., Rosconi, F., Rui, Z., Tian, Z., Tobias, N.J., Tsunematsu, Y., Wiemann, P., Wyckoff, E., Yan, X., Yim, G., Yu, F., Xie, Y., Aigle, B., Apel, A.K., Balibar, C.J., Balskus, E.P., Barona-Gómez, F., Bechthold, A., Bode, H.B., Borriss, R., Brady, S.F., Brakhage, A.A., Caffrey, P., Cheng, Y.Q., Clardy, J., Cox, R.J., De Mot, R., Donadio, S., Donia, M.S., Van Der Donk, W.A., Dorrestein, P.C., Doyle, S., Driessen, A.J.M., Ehling-Schulz, M., Entian, K.D., Fischbach, M.A., Gerwick, L., Gerwick, W.H., Gross, H., Gust, B., Hertweck, C., Höfte, M., Jensen, S.E., Ju, J., Katz, L., Kaysser, L., Klassen, J.L., Keller, N.P., Kormanec, J., Kuipers, O.P., Kuzuyama, T., Kyrpides, N.C., Kwon, H.J., Lautru, S., Lavigne, R., Lee, C.Y., Linquan, B., Liu, X., Liu, W., Luzhetskyy, A., Mahmud, T., Mast, Y., Méndez, C., Metsä-Ketelä, M., Micklefield, J., Mitchell, D.A., Moore, B.S., Moreira, L.M., Müller, R., Neilan, B.A., Nett, M., Nielsen, J., O'Gara, F., Oikawa, H., Osbourn, A., Osburne, M.S., Ostash, B., Payne, S.M., Pernodet, J.L., Petricek, M., Piel, J., Ploux, O., Raaijmakers, J.M., Salas, J.A., Schmitt, E.K., Scott, B., Seipke, R.F., Shen, B., Sherman, D.H., Sivonen, K., Smanski, M.J., Sosio, M., Stegmann, E., Süssmuth, R.D., Tahlan, K., Thomas, C.M., Tang, Y., Truman, A.W., Viaud, M., Walton, J.D., Walsh, C.T., Weber, T., Van Wezel, G.P., Wilkinson, B., Willey, J.M., Wohlleben, W., Wright, G.D., Ziemert, N., Zhang, C., Zotchev, S.B., Breitling, R., Takano, E., Glöckner, F.O.: Minimum Information about a Biosynthetic Gene cluster. Nature Chemical Biology **11**(9), 625–631 (2015). doi:10.1038/nchembio.1890
25. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L.: BLAST+: architecture and applications. BMC Bioinformatics **10**(1), 421 (2009). doi:10.1186/1471-2105-10-421
26. Huerta-Cepas, J., Serra, F., Bork, P.: ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Molecular Biology and Evolution **33**(6), 1635–1638 (2016). doi:10.1093/molbev/msw046
27. Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.Y., Lopez, R., Hunter, S.: InterProScan 5: Genome-scale protein function classification. Bioinformatics **30**(9), 1236–1240 (2014). doi:10.1093/bioinformatics/btu031
28. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., De Hoon, M.J.L.: Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics **25**(11), 1422–1423 (2009). doi:10.1093/bioinformatics/btp163. arXiv:1011.1669v3
29. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G.: Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology **7**(1), 539 (2011). doi:10.1038/msb.2011.75. 0-387-31073-8
30. Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T.: trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics **25**(15), 1972–1973 (2009). doi:10.1093/bioinformatics/btp348
31. Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q.: IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution **32**(1), 268–274 (2015). doi:10.1093/molbev/msu300
32. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., Jermiin, L.S.: ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates. Nature Methods **14**(6), 587–591 (2017). doi:10.1038/nmeth.4285
33. Minh, B.Q., Nguyen, M.A.T., Von Haeseler, A.: Ultrafast approximation for phylogenetic bootstrap. Molecular Biology and Evolution **30**(5), 1188–1195 (2013). doi:10.1093/molbev/mst024
34. Paradis, E., Claude, J., Strimmer, K.: APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics (Oxford, England) **20**(2), 289–90 (2004)
35. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2017). R Foundation for Statistical Computing. https://www.r-project.org/
36. Wickham, H.: ggplot2: Elegant Graphics for Data Analysis. Springer, ??? (2009). http://ggplot2.org
37. Yu, G., Smith, D., Zhu, H., Guan, Y., Lam, T.T.-Y.: ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution (2017). doi:10.1111/2041-210X.12628
38. Henry, L., Wickham, H., Chang, W.: ggstance: Horizontal 'ggplot2' Components (2016)
39. Hunter, J.D.: Matplotlib: A 2D graphics environment. Computing In Science & Engineering **9**(3), 90–95 (2007). doi:10.1109/MCSE.2007.55
40. Sullivan, M.J., Petty, N.K., Beatson, S.A.: Easyfig: A genome comparison visualizer. Bioinformatics **27**(7), 1009–1010 (2011). doi:10.1093/bioinformatics/btr039

Figures
Tables

Table 1 Number of hybrids and total SM proteins in *Aspergillus* species and *Penicillium chrysogenum*. SM proteins comprise PKS, PKS-like, NRPS, NRPS-like, Hybrids, dimethylallyltryptophan synthase and terpene cyclases.

Complete name	Jgi name	Section/Group	hybrids	Total SM proteins
A. campestris	Aspcam1	*Candidi*	5	48
A. steynii	Aspste1	*Circumdati*	10	98
A. clavatus	Aspcl1	*Clavati*	4	43
A. flavus	Aspfl1	*Flavi*	3	72
A. oryzae	Aspor1	*Flavi*	2	68
A. fumigatus A1163	Aspfu_A1163_1	*Fumigati*	2	36
A. fumigatus Af293	Aspfu1	*Fumigati*	1	37
A. novofumigatus	Aspnov1	*Fumigati*	4	59
A. nidulans	Aspnid1	*Nidulantes*	1	63
A. brasiliensis	Aspbr1	*Niger_biseriates*	5	82
A. carbonarius	Aspca3	*Niger_biseriates*	5	65
A. costaricaensis	Aspcos1	*Niger_biseriates*	5	97
A. ellipticus	Aspell1	*Niger_biseriates*	2	72
A. eucalypticola	Aspeuc1	*Niger_biseriates*	6	78
A. heteromorphus	Asphet1	*Niger_biseriates*	3	58
A. ibericus	Aspibe1	*Niger_biseriates*	6	57
A. luchuensis CBS 106.47	Aspfo1	*Niger_biseriates*	6	93
A. luchuensis IFO 4308	Aspka1_1	*Niger_biseriates*	6	88
A. neoniger	Aspneo1	*Niger_biseriates*	5	82
A. niger ATCC 1015	Aspni7	*Niger_biseriates*	9	86
A. phoenicis	Aspph1	*Niger_biseriates*	9	89
A. piperis	Asppip1	*Niger_biseriates*	6	86
A. sclerotiicarbonarius	Aspscle1	*Niger_biseriates*	6	82
A. sclerotioniger	Aspscl1	*Niger_biseriates*	5	75
A. tubingensis	Asptu1	*Niger_biseriates*	7	93
A. vadensis	Aspvad1	*Niger_biseriates*	6	85
A. welwitschiae	Aspwel1	*Niger_biseriates*	7	87
A. aculeatinus	Aspacu1	*Niger_uniseriates*	3	90
A. aculeatus	Aspac1	*Niger_uniseriates*	4	73
A. brunneoviolaceus	Aspbru1	*Niger_uniseriates*	2	85
A. fijiensis	Aspfij1	*Niger_uniseriates*	4	94
A. homomorphus	Asphom1	*Niger_uniseriates*	3	78
A. indologenus	Aspind1	*Niger_uniseriates*	6	92
A. saccharolyticus	Aspsac1	*Niger_uniseriates*	2	52
A. uvarum	Aspuva1	*Niger_uniseriates*	3	76
A. violaceofuscus	Aspvio1	*Niger_uniseriates*	5	79
A. ochraceoroseus	Aspoch1	*Ochraceorosei*	2	30
P. chrysogenum	Pench1	*Penicillium*	2	50
A. terreus	Aspte1	*Terrei*	1	73

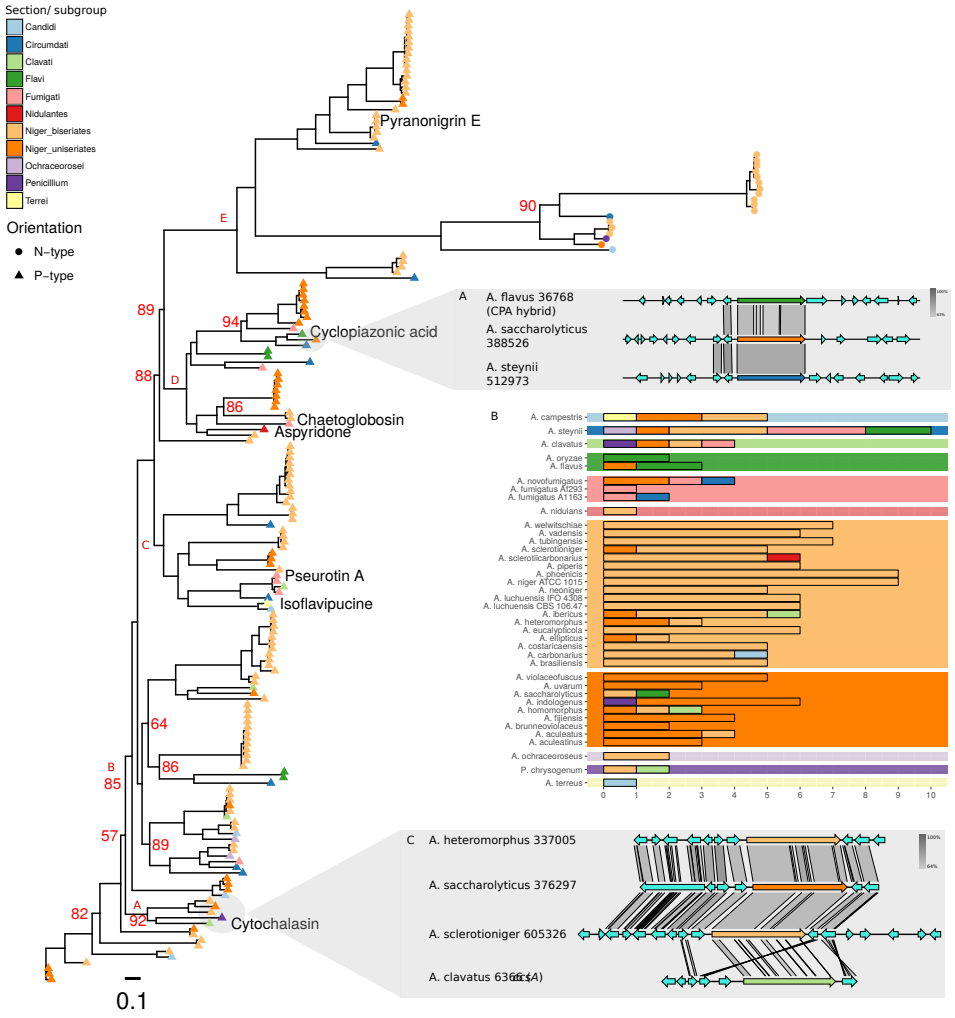


Figure 1 Hybrid dynamics throughout Aspergilli ML phylogeny of PKS-NRPS and NRPS-PKS hybrid proteins was created on aligned and trimmed protein sequences. Red letters indicate subtrees shown in Fig. S1. Sections and species groups indicated by tip color; Orientation of hybrids N-type (NRPS-PKS) and P-type (PKS-NRPS) indicated by tip shape. (**A**) Synteny plots of cyclopiazonic acid related hybrids. (**B**) Classification of nearest neighbors of ML phylogeny. A matrix of tip distances was extracted from the tree and nearest neighbors classified according to their section. (**C**) Synteny plot of cytochalasin related hybrids (ccsA in *A. clavatus*). *A. sclerotioniger* is known to produce sclerotionigrin — a cytochalasan

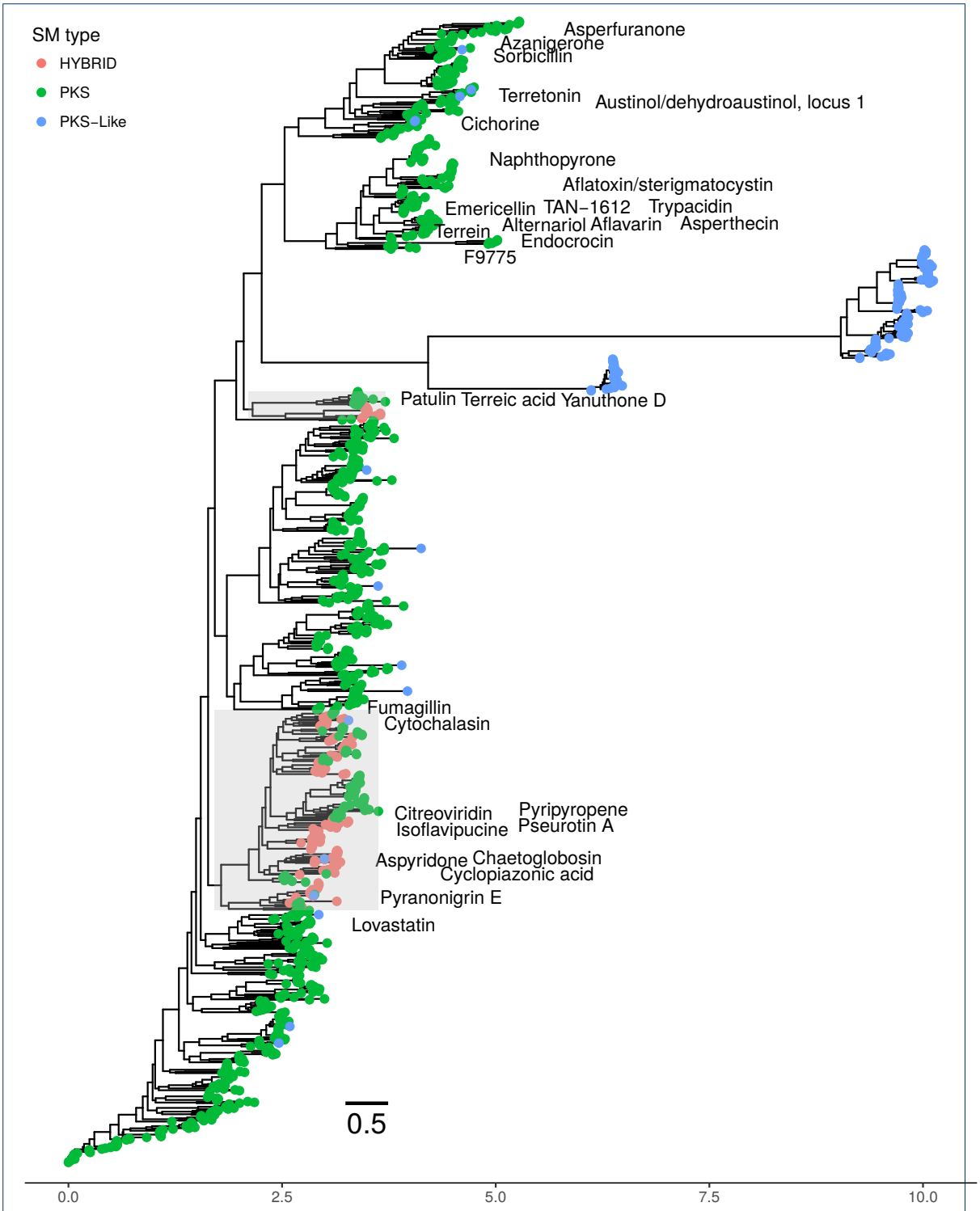


Figure 2 Phylogeny of PKS, PKS-like and hybrid proteins. The maximum likelihood phylogeny was created from KS domains of PKS, PKS-like and hybrid proteins. Tip color shows SM gene type (red: hybrid, green: PKS, blue: PKS-like). Hybrids linked to compounds are labelled with the compound name. Grey boxes highlight groups containing hybrids.

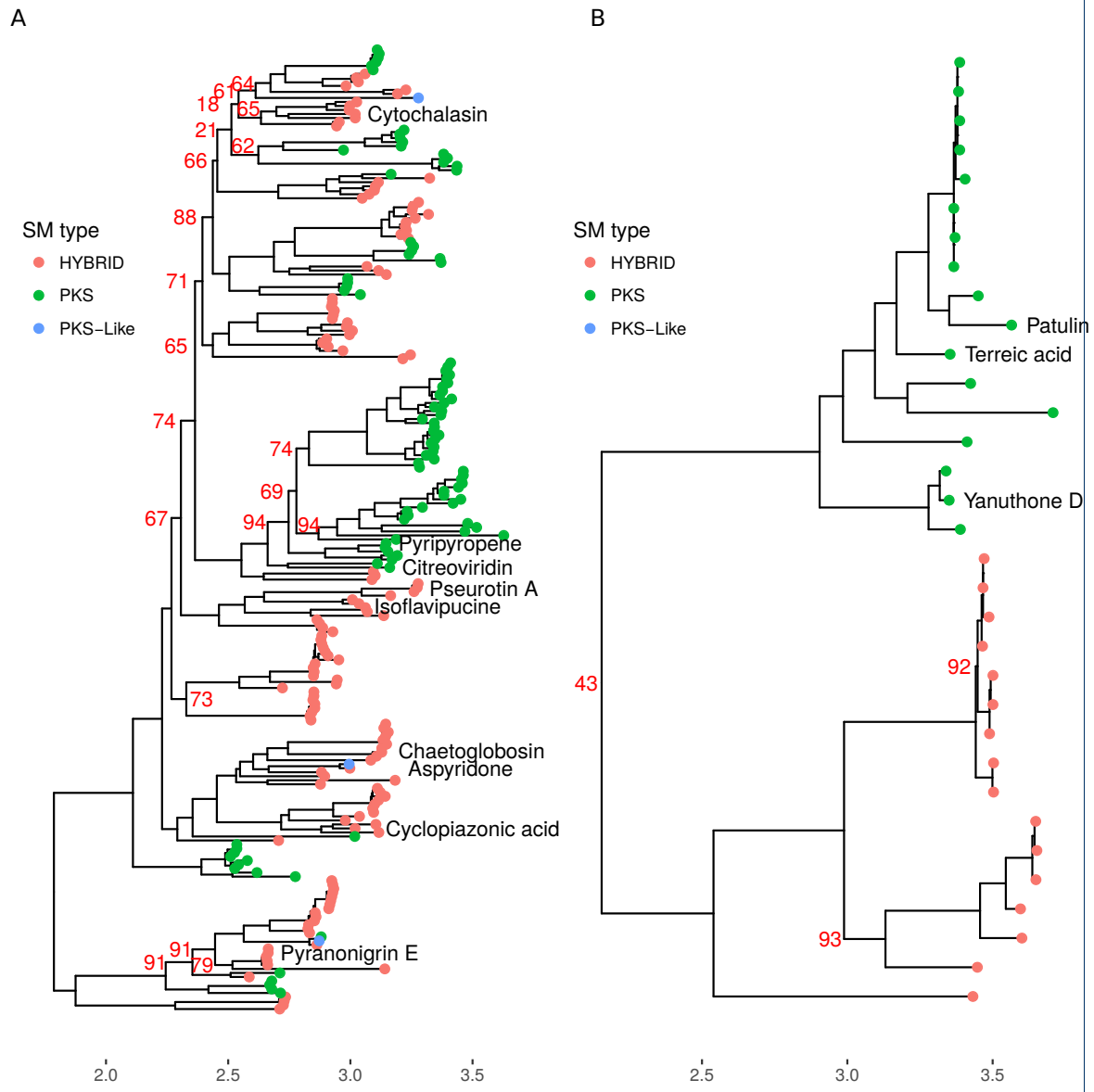


Figure 3 Phylogeny of PKS, PKS-like and hybrid proteins Left: PKS-NRPS containing subtree; Right: NRPS-PKS containing subtree. The maximum likelihood phylogeny was created from KS domains of PKS, PKS-like and hybrid proteins. Tip color shows SM gene type (red: hybrid, green: PKS, blue: PKS-like). Tip labels show associated compound, if applicable. Bootstrap values under 95 are shown in red. Hybrids linked to compounds are labelled with the compound name.

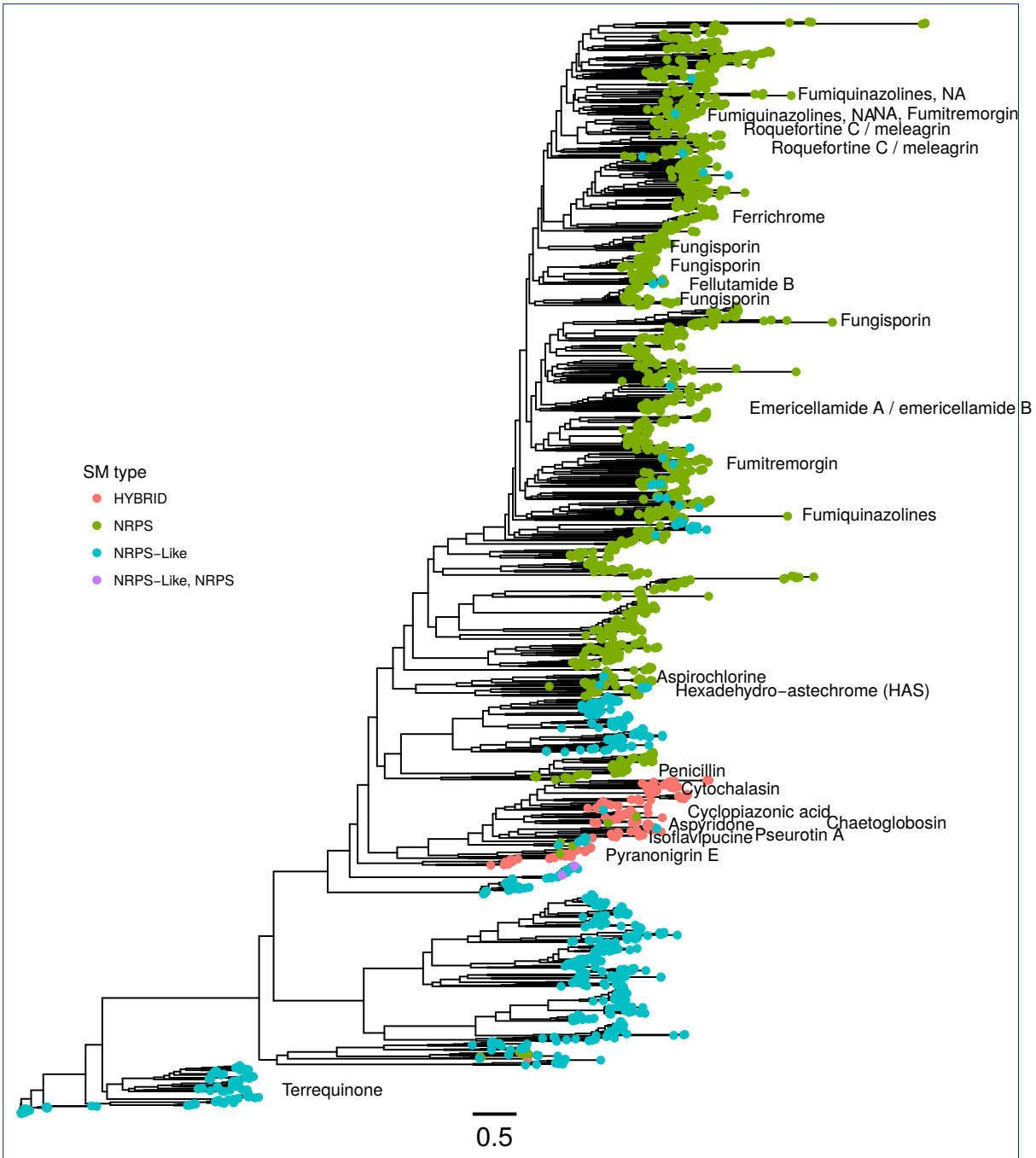


Figure 4 ML phylogeny of A domains. Tip colors indicate SM protein type; Tip labels show associated compound (if applicable). The phylogenetic tree shows that hybrids are monophyletic when compared to NRPSs and NRPS-likes.

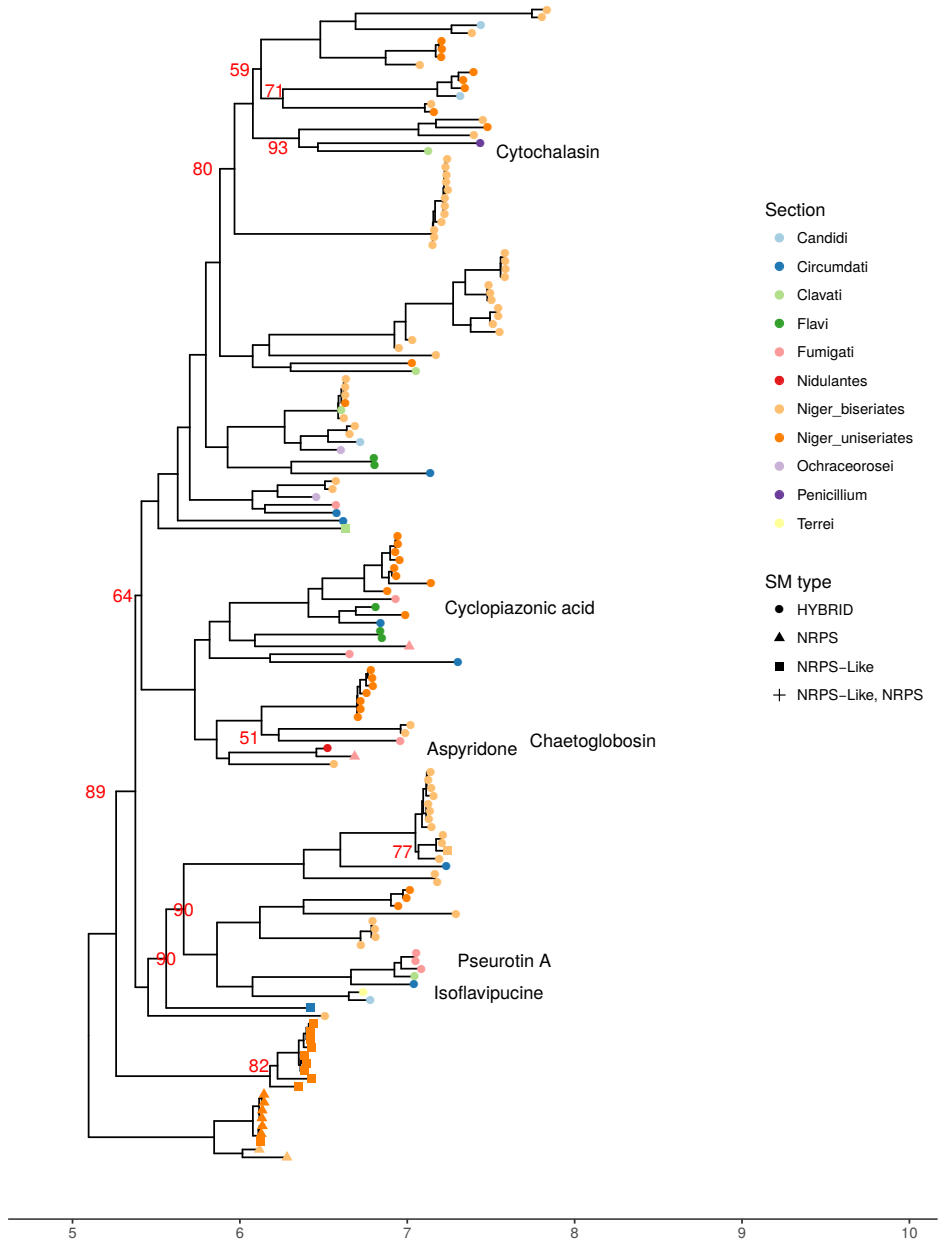


Figure 5 Subtree of ML phylogeny of A domains Tip color indicates section/subgroup; tip shape indicates SM protein type; tip label shows associated compound (if applicable). A group of NRPS and NRPS-likes from uniseriate *Nigri* species forms a sister clade to the monophyletic hybrids. We only find two NRPSs of *Fumigati* species and one of *A. steynii* (section *Circumdati*) inside hybrid groups.

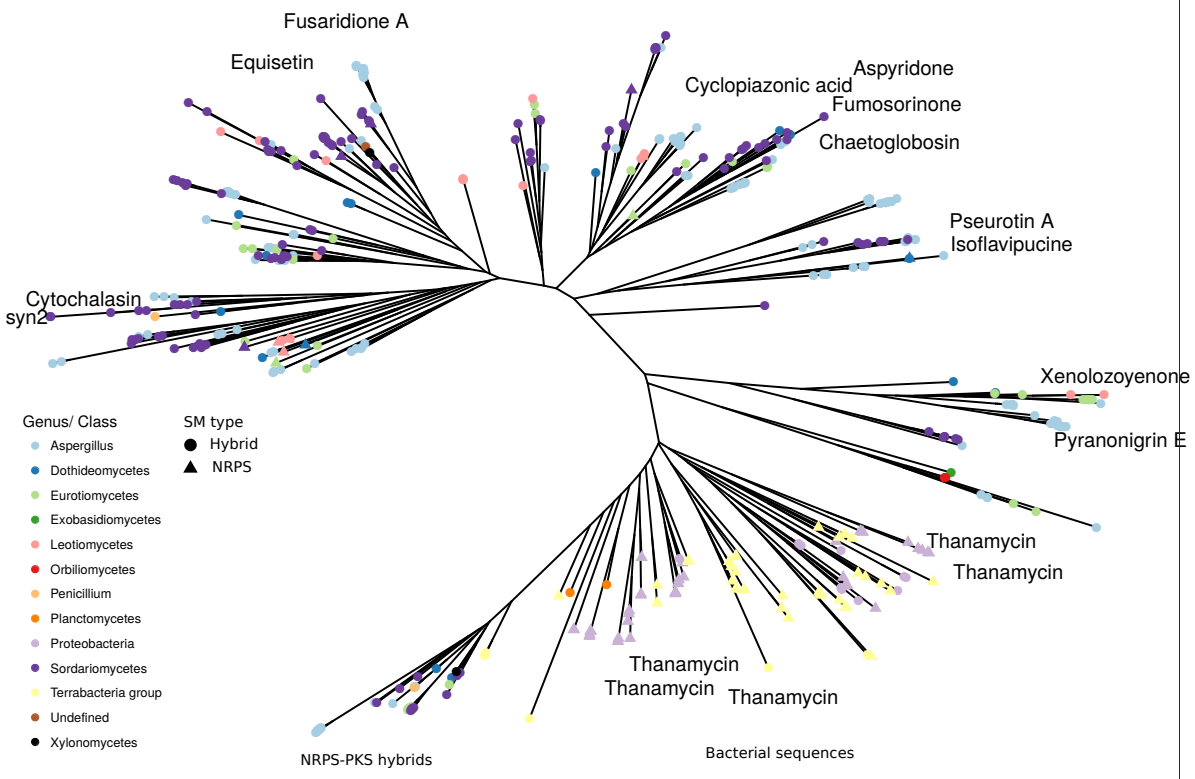


Figure 6 ML phylogeny of A domains from PKS-NRPS and NRPS-PKS hybrid genes and their best blast hits from the NCBI non-redundant protein database. Protein sequences were aligned with clustal omega, trimmed and filtered using trimal, and a ML phylogeny generated using iqtree. Compounds of hybrids, if present in MIBiG are shown above tips. Thanamycin is shown multiple times because each domain of the lipopeptide synthetase is annotated. Tip colors indicate whether the hybrid domain is derived from the genera *Aspergillus*, *Penicillium* or a different taxonomical class.

Chapter 5

Comparative genomics of *A. nidulans* and section *Nidulantes*

Aspergillus nidulans has long served as the sole representative of its section in comparative genomics studies. In this study we are *de novo* sequencing section *Nidulantes* and using our established pipeline to investigate whether proteins and secondary metabolites of *A. nidulans* are generally present in the whole section. This is an important step in our effort to sequence the whole genus, since it will provide an estimate of how applicable the conclusions of a model organism are to the whole section. Furthermore, we investigate general regulators and their distribution throughout the section to get insights into fungal speciation. Copy number variations as well as gene loss of central regulators and developmental genes has been observed. We also find differences in central pathways like pigmentation. The chapter will give an insight into the distribution of investigated proteins throughout section *Nidulantes* and other *Aspergillus* species and identify whether proteins are essential or can be replaced by paralogs. Additionally, we will investigate the dynamics of secondary metabolite gene clusters and potential horizontal gene transfers.

Comparative genomics of *Aspergillus nidulans* and section *Nidulantes*

Sebastian Theobald[a], Tammi Vesth[a], Jane Lind Nybo[a], Inge Kjæbølling[a], Robert Riley[b], Asaf Salamov[b], Ellen Kirstine Kyhne[a], Martin Engelhard Kogle[a], Jens Christian Frisvad[a], Jakob Blæsbjerg Hoof[a], Uffe H. Mortensen[a], Paul Dyer[c], Michelle Momany[e], Thomas Ostenfeld Larsen[a], Scott Baker[d], and Mikael Rørdam Andersen[a,1]

[a]Department of Biotechnology and Biomedicine, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark; [b]Department of Energy Joint Genome Institute, Walnut Creek, CA 94598; [c]School of Life Sciences, University of Nottingham, Nottingham, UK NG7 2RD; [d]Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, USA; [e]Fungal Biology Group & Plant Biology Department, University of Georgia,Athens, Georgia, USA 30602

This manuscript was compiled on February 1, 2018

***Aspergillus nidulans* is an important model organism for eukaryotic biology and has served as the reference for the section *Nidulantes* in comparative genomics studies. In this study, we *de novo* sequenced 25 species. Whole-genome phylogeny of 34 *Aspergillus* species and *Penicillium chrysogenum* clarifies the position of clades inside section *Nidulantes*. Comparative genomics reveals a high genetic diversity between species with 684 up to 2433 unique protein families. Nonetheless, protein families for investigated general regulators and developmental proteins are conserved. Furthermore, we categorized 2118 secondary metabolite gene clusters (SMGC) into 603 families across Aspergilli, with at least 40% of the families shared between *Nidulantes* species. Genetic dereplication of SMGC and subsequent synteny analysis provides evidence for horizontal gene transfer of a SMGC. Our analysis shows that proteins investigated in *A. nidulans* as well as its SMGC families are generally present in the section *Nidulantes*, supporting its role as model organism.**

Aspergillus | Comparative genomics | Secondary metabolism

Research on *Aspergillus nidulans* has benefited the scientific community in many ways. Its use as a model organism for eukaryotic biology has provided insights into cell polarity and cell cycle mechanisms (1), DNA repair mechanisms (2), morphogenesis (3) and cytosceleton development (4). In addition, studies on *A. nidulans* has added important information to the understanding of antifungal drug resistance (5–7).

A. nidulans is part of section *Nidulantes* which has been synonymized with section *Versicolores* and *Aenei* — defining them as subclades of section *Nidulantes* (8). Members of the section are mainly decomposers of plant material, but also include species associated with human infections in the case of *A. unguis* and *A. versicolor* — the latter affecting indoor environments (9) while also being used as source for xylanases (10). The section also includes the coral pathogen (11) *A. sydowii*.

Characteristic to species of the section are, beside the morphological characters, their secondary metabolites (SMs). Sterigmatocystin — a carcinogen (12) and common food contaminant (13) — although produced throughout many subgenera, is mostly present in Aspergilli of section *Nidulantes* and *Circumdati* (14). Besides mycotoxins, some species also produce medically relevant SMs, as e.g. penicillin (15–17). Large interest in SMs has driven the investigation of secondary metabolite gene clusters (SMGCs) (18–20).

Although *A. nidulans* has been used extensivly as a model organism the question of whether it is an appropriate model has remained to be investigated. Sequencing of *A. nidulans* and comparisons to *A. fumigatus* and *A. oryzae* (21) have highlighted similarities between these fungi, but have not covered sufficient species to proof the general applicability of *A. nidulans* as a model. We are the first ones to set *A. nidulans* into genomic context to its close relatives of section *Nidulantes*.

In this study, we investigate the dynamics of selected homologous protein families throughout 34 *Aspergillus* (and *P. chrysogenum*) species to assess the application of *A. nidulans* as a model organism for these species. We provide further insights into the phylogeny of section *Nidulantes* using whole-genome-phylogenetic approaches and investigate the diversity of secondary metabolite gene clusters (SMGCs).

Results

Whole genome phylogeny of section *Nidulantes*. Phylogenetic analysis provides insight into the taxonomy of species. Multiple genes were used to improve taxonomy of section Nidulantes but still disagreed on section boundaries (8, 22). With the increasing number of genome sequences available, we sought to improve the section definitions and species relationships using whole-genome phylogenetic methods.

We constructed a whole genome phylogeny of the dataset using 200 concatenated genes to increase the resolution of the maximum likelihood (ML) phylogeny of section *Nidulantes* and confirm relationships among clades. The ML phylogeny shows high bootstrap support for most species and is hence providing consistent results based on our gene selection. Phylogenetic distance was used to establish clades among clustering species and bootstraps were used to support the hypothesis of clade separation (Fig. 1).

Using the ML tree we can identify several clades within section *Nidulantes*. The *A. nidulans* clade (clade I) shows a clear distinction from other species of section *Nidulantes* and can be further subdivided into three groups. Species around *A. nidulans* are forming one subgroup. The residual species of clade I are included in a second subgroup with only *A. multicolor* forming a third group. Our method resolves all branches with high support, except for internal branches for *A. indicus*, *A. similis* and *A. navahoensis* which shows bootstrap support of 79% and 76%, respectively. Following is *A. unguis* with 68%. The *A. stella-maris* clade (clade II) shows clear separation from other species. *A. undulatus* forms the outgroup of this clade. The *A. versicolor* clade (clade III) is placed as a sister clade to the *A. aurantiobrunneus*

The authors declare no conflict of interest.

[2]To whom correspondence should be addressed. E-mail: mrbio.dtu.dk

and *A. varians*. The most phylogenetically distant clade is the *A. aeneus clade* (clade VI) with species *A. spectabilis*, *A. crustosus* and *A. karnatakaensis*. The separation of section *Nigri* and *Nidulantes* only shows 79% bootstrap support.

Categorizations by Hubka et al (2016) (8) and Chen et al (2016) (22) using genetic methods and a polyphasic approach identified similar clades of section *Nidulantes*. They used (coding regions of) *benA*, the calmodulin gene (*caM*), RNA polymerase II gene (*RPB2*) and internal transcribed spacer (ITS) sequences to establish a phylogeny of species using ML phylogeny.

Our method shows that whole-genome phylogenies can be used to separate section *Nidulantes* into clades of closely related species with high bootstrap support. The phylogenetic distance of *A. undulatus* in regard to clade II is relatively large which suggests a separation into its own clade. Furthermore, bootstrap values under 80% could be a result of the increased resolution of our phylogeny. Our results are in accordance with previous studies based on single gene trees regarding the general organization of section *Nidulantes*.

Genome statistics. Genome sizes in the species of *Nidulantes* range from 26.1 to 38.7 Mb, with a GC content of 45.5–50.8%. The number of predicted proteins range from 8924 to 13620. Clade IV tends to have larger genomes than other *Nidulantes* species (37.5 Mb on average), while *A. unguis* shows a smaller genome size (26.1 Mb). These statistics are similar to the overall *Aspergillus* genomes reported previously (26). Only *A. unguis* and clade IV deviate from the general Aspergilli genome sizes.

Core pan proteome. To investigate section *Nidulantes* proteome diversity and *A. nidulans* similarity to other species, we generated homologous protein families across all species and related them to the phylogenetic tree (Fig. 2).

The protein families found in all members of the Aspergilli (*Aspergillus* core proteome) covers 4033 families. *Nidulantes* species share 105 protein families (Nidulantes specific families). Remarkably, species up until clade I only share very few core protein families. For clade I we can identify 39 clade-specific protein families which increases to 44 for the two clade I subgroups. *A. multicolor* contains 2411 species-unique protein families, which is quite surprising considering that we included many species of its own clade. In comparison, clade IV shows approximately 2400 species-unique protein families per species as well — only by including three species from the same clade. Thus, emphasizing the distance from *A. multicolor* to clade I. 446 protein families are unique to *A. fructiculosus* and *A. falconensis* suggesting them to be isolates of the same species — which is further supported by the whole-genome phylogeny (Fig. 1). *A. nidulans* only contains 682 protein families which are species-unique, suggesting that much of its proteome is representative of section *Nidulantes*.

The protein families of section *Nidulantes* are distinct from the ones observed earlier for section *Nigri* since the number of isolates and the phylogenetic distance of species are differnt (27). However, we can show that in both cases species show around 1000 species-unique proteins, this similarity drops to under 400 species-unique proteins when isolates of the same species are compared. Section *Nidulantes* tends to share more section-unique proteins than clade-unique proteins. *Nigri* species tend to share more proteins on the clade-unique level. This indicates that the analyzed species in section *Nidulantes* are overall more conserved as a section than section *Nigri*.

Dynamics of general regulators throughout Aspergilli. The discovery of general regulators in *A. nidulans* led to a better understanding of e.g. carbon catabolite repression through CreA (28), or pH response through PacC (29), or regulation of secondary metabolism LaeA (30) (among others). Using the protein families described above, we investigated the diversity of general regulators and whether they are part of the core or accessory proteome.

We identified protein families with one protein per species for CreA, (31); PacC, a regulator of transcription responding to alkaline pH (29); and McrA, a master regulator of SM (32). LaeA, another regulator of SM (30), is found in a core family with an additional copy in *A. unguis*, *A. foveolata* and *A. aculeatinus*. ML phylogeny of predicted LaeA homologs shows that the second copy in *A. foveolata* is from clade II (Fig. S6). *A. aculeatinus* seems to contain a *laeA* homolog from *A. terreus* and the second copy in *A. unguis* has no apparent close relative in this set of species.

The regulator GalR is unique to *A. nidulans* according to (33). We identified a protein family containing both GalR and AraR (Fig. S5). ML phylogeny of homologs shows AraR and GalR in separate branches in section *Nidulantes* and *A. ochraceroseus* with high bootstrap support. Thus GalR is present in all species of section *Nidulantes* and potentially in *A. ochraceroseus*. WscB, a putative stress sensor (34), is unique to species of clade I, while WscA is part of the core proteome. Absence of one part of the complex suggests that WscB is more variable than WscA.

In summary, we are able to find important general regulators in the core proteome, emphasizing that investigated proteins in *A. nidulans* are conserved throughout species. McrA in other *Aspergillus* species is a good target for overexpression studies. Previous studies have shown that McrA overexpression induces production of secondary metabolites (35). LaeA shows additional copies in some species which, according to their phylogeny, seem to be paralogs of different species (Fig. S6). Hence we suggest that LaeA is a specialized rather than a global regulator (30). GalR, which was suggested to be unique to *A. nidulans* (33) is predicted to have homologs in all *Nidulantes* species. Whether the *A. ochraceroseas* copy is a GalR homolog could not be identified on the sole basis of ML phylogeny (Fig. S5). If it is a GalR then this regulator is distributed through the whole subgenus *Nidulantes*.

Analysis of proteins involved in polarity, cytoskeleton and cell cycle. The highly polar growth of filamentous fungi makes them ideal for studies of morphology and its coordination with the cell cycle. The spatial separation of nuclei and well-developed genetic system of *A. nidulans* has made it an attractive model for many of these studies (1, 36, 37).

To investigate whether *A. nidulans* is a good model for morphology and cell cycle in other Aspergilli, we identified homologs of key polarity genes including the Rho GTPase Cdc42 (modA), its associated guanine nucleotide exchange factor (GEF) Cdc24 and GTPase-activating protein (GAP) Rga1, and the p21-activated kinase Cla4. We also identified homologs of cytoskeletal genes including actin, tubulins and septins and key cell cycle genes including cyclin dependent kinases, wee1 kinase, phosphatases and anaphase promoting

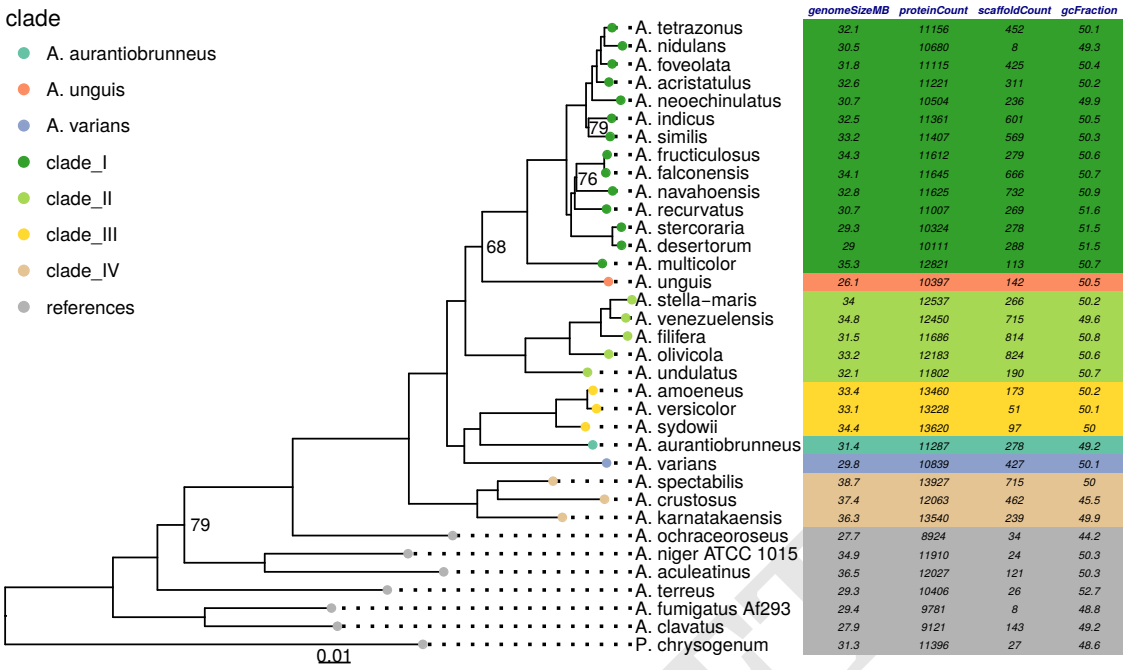


Fig. 1. Dendrogram of species. Whole genome phylogeny constructed from alignment of 200 bidirectional best hits between species using RAxML (23), MAFFT (24), and Gblocks (25). Bootstrap support shown at node if under 80. Tip color indicates clade. Genome statistics shown in right panel.

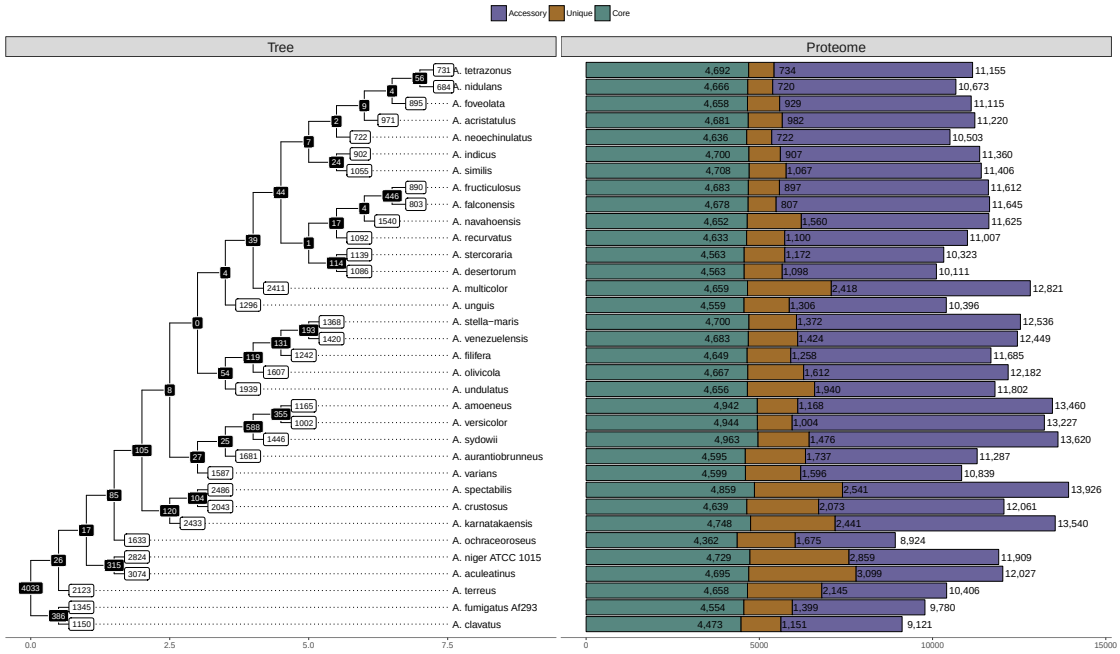


Fig. 2. Core, accessory and unique proteins through *Aspergillus* species. Protein families were built using single linkage on bidirectional protein blast hits with a percent identity of at least 50% and sum of coverage (query and subject) of at least 130% . (Left) Counts on nodes show unique protein families for species included in the branch; counts on tips show unique protein families for individual organisms. (Right) Core, unique and accessory proteins represented in stacked bars per species.

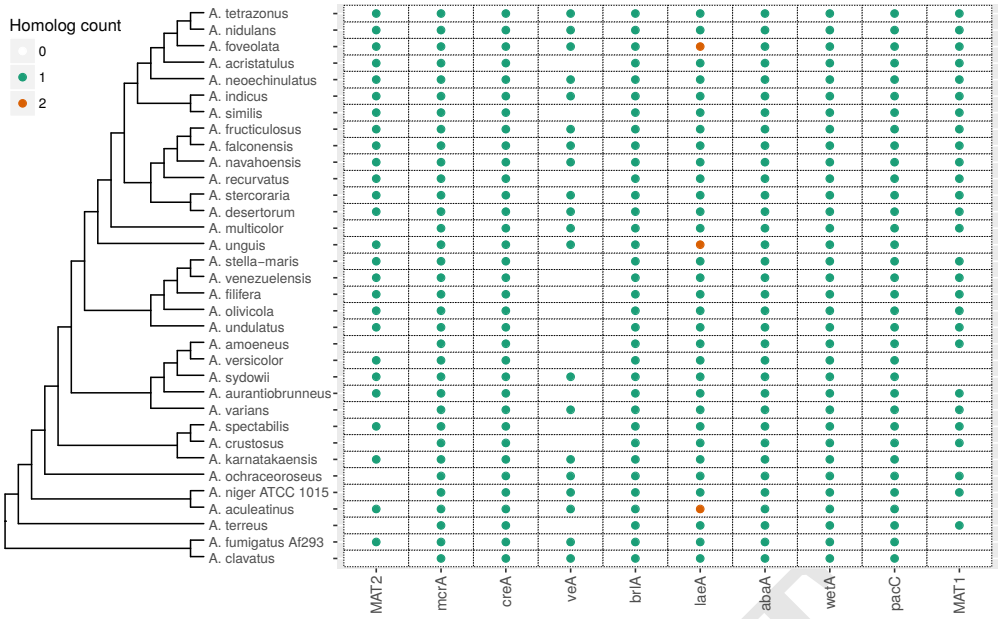


Fig. 3. Selection of protein families with known examples in dataset. Protein families are shown as columns with a dots indicating copy number by color. Column labels correspond to annotated *A. nidulans* genes.

complex (APC) components (Fig. 4). As expected, multiple Cdc42 homologs were present in all Aspergilli examined and included Cdc42, RacA, Rho4, and RhoA along with 2-3 other uncharacterized Rho GTPases. Single homologs of Rga1 and Cla4 were also present in all species examined. While most of the species examined had a single Cdc24 GEF homologue, five of the clade III and clade IV *Aspergillus* species lacked a Cdc24 GEF. This is surprising because Cdc24 is an essential gene in *A. nidulans* (38). The lack of Cdc24 homologs in five of 34 species examined, along with conservation of other key polarity genes, suggests that a more highly diverged gene might encode a GEF for Cdc42 in these species.

Single homologs of polarisome components SpaA, BudA, BemA and Ste20 were present in all species, while four lacked the formin SepA (39). All species contained single or multiple homologs of cytoskeletal elements actin, actin related protein Arp1 (nudK) and tubulins. Alpha and beta tubulins ranged from 1-3 homologs per species, consistent with previously reported presence of tubulins specialized for specific developmental states (40, 41) The core septins AspA^{CDC11}, AspB^{CDC3}, AspC^{CDC12}, AspD^{CDC10} had single homologs in each species, consistent with the assembly of these core septins into complexes (42). Strikingly the noncore septin AspE was absent in 11 of the 34 species examined. This is similar to the patchy distribution of this nonessential septin across fungi and across kingdoms (42–44). All of the key cell cycle genes examined had at least one homologue present in all species with the exception of the wee1 kinase AnkA which was absent in two clade I species. AnkA is an essential gene in *A. nidulans* (45), so its absence in some species is surprising and once more suggests that a gene too highly diverged to be detected might play the same role in these species.

In terms of general trends, we found that extra homologs of polarity, cytoskeletal and cell cycle genes were common across the *Aspergillus* species examined, with all 34 species examined showing the gain of 1-3 homologs. Complete loss of all homologs of a specific gene was less common, occurring in 16/34 species examined, with the nonessential septin AspE accounting for half of the cases. Loss of a gene family was most common in clades III and IV. *A. nidulans* contained representatives of each polarity, cytoskeletal, or cell cycle gene examined and so in terms of gene content is a good model for all of the *Aspergillus* species examined.

Analysis of CAZymes. Carbohydrate active enzymes (CAZymes) determine the plant biomass degrading potential of fungi. Aspergilli contain many of these CAZymes which enables them to degrade numerous polysaccharides (46).

Throughout clades, the *Aspergillus* species of section *Nidulantes* show generally the same diversity of CAZymes while differing in the amounts of specific classes 5. In the auxiallary category aryl alcohol oxidase and glucose 1-oxidase (AA3_2) are the predominant classes, followed by copper-dependent lytic polysaccharide monooxygenases (AA9). Carbohydrate binding modules (CBM) show two dominant classes in *Nidulantes*: CBM50, CBM18; a contrast to *Aspergillus* references where CBM1 is the most abundant class, followed by CB50 and CBM18. Glycosylhydrolases (GH) GH3, GH43 and GH18 are the most common ones in *Nidulantes* species, except for *A. aurantiobrunneus* which contains slightly more GH16 than GH18. GH28 is increased in *Aspergillus references* compared to *Nidulantes*. Glycosyltransferases (GT) and polysaccharide lyases (PL) show similar abundance patterns through *Aspergillus* species. The varying amounts of CAZyme classes indicates the specialization of clades towards certain polysaccharides.

Secondary metabolite gene cluster diversity confirms clade concepts. Aspergilli show a vast diversity of secondary metabolites (SMs), with dynamics in secondary metabolite gene clusters (SMGCs) throughout species leading to new compounds.

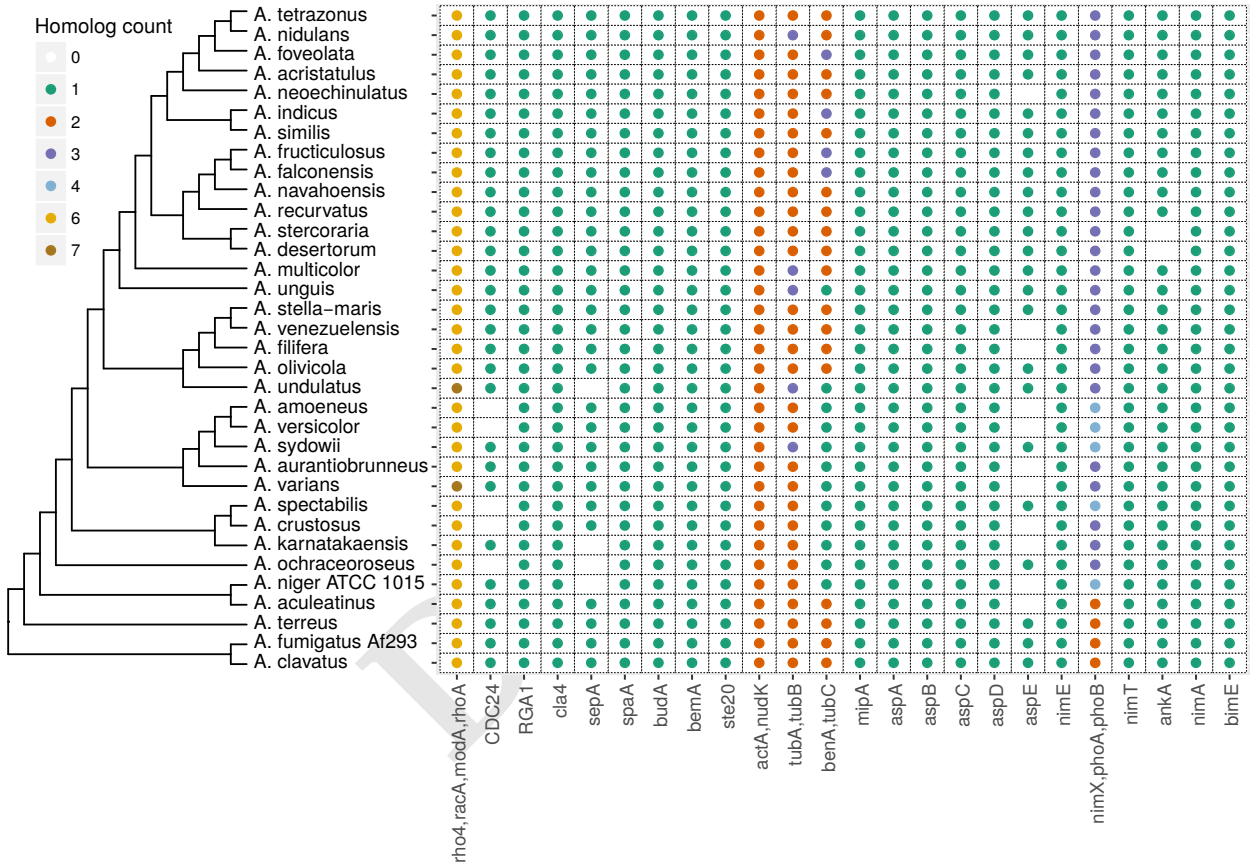


Fig. 4. Distribution of polarity, cytoskeleton and cell cycle proteins throughout the dataset. Protein families for *A. nidulans* proteins (one family per column) involved in polarity, cytoskeleton and cell cycle were extracted from protein families and annotations concatenated if multiple annotated proteins were found per family (annotated on the x-axis). Homologs are shown for all species of the dataset. Number of homologs shown by point color.

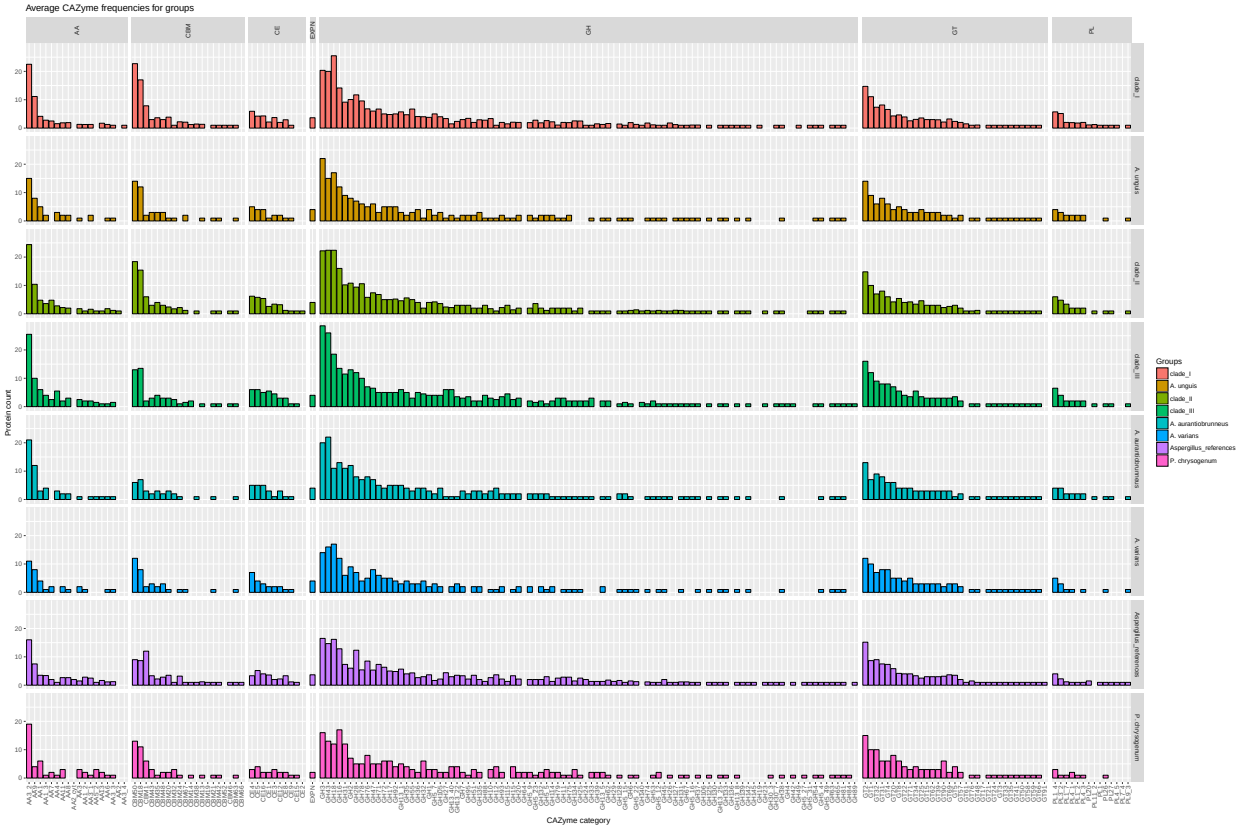


Fig. 5. Analysis of CAZymes. Bars show the average number of proteins in each clade (references split up and averaged). Group indicated by color. Top panels show categories: Auxiliary activities (AA), carbohydrate-binding molecules (CBM), carbohydrate esterases (CE), Distant plant expainins (EXPN), glycoside hydrolases (GH), Glycosyltransferases (GT), polysaccharide lyases (PL)

Hence, it was an obvious step to aggregate SMGCs into homologous families to describe the variety of SMGCs through species in section *Nidulantes* and investigate whether they further confirm the assignment of species into clades.

We compared all SMGC members and calculated a similarity score for each SMGC pair. Using random walk clustering on resulting similarity networks of SMGCs categorized 2118 SMGC into 603 families. Of these families 433 are unique to a species, 175 of the unique gene clusters are belonging to references (Supplementary Fig. 1). 29 SMGC families are shared between two species, while 89 are shared between 3-10 species. Identifying the shared secondary metabolite gene cluster (SMGC) families throughout the section *Nidulantes*, we can determine that species share 40-50% of their SMGC families (Fig. 6). Clustering the species by their shared SMGC content, we can identify subgroups inside the section which are to a large extent in agreement with clades based on phylogeny.

Species inside clade I share at least 50% of their SMGC families, while many members in subgroup I share 70-80% SMGC families, meaning that their repertoire of SMs is largely the same. *A. fructiculosus* and *A. falconensis* in subgroup II share between 90-100% of their SMGC families, an amount commonly shared by isolates of the same species. *A. multicolor* shares the lowest amount of SMGC with the whole its clade (40-60%). *A. unguis* shares mostly 40-50% SMGC with clade I and 30-40% with the rest of the section — confirming that this species forms its own clade. Species in clade III show a high conservation of their SMGC content with 70-80% SMGC families shared between species. An exception is *A. undulatus*, a member of clade III, which only shares 50-60% SMGC families with the other members of it's own clade — also indicating it as its own clade. The SMGC similarity drops with *A. aurantiobrunneus*, *A. varians* and clade IV — only showing low conservation in SMGC families. This suggests a different SM content in these species than in other *Nidulantes* species.

In summary, we can further sustain the concept of section *Nidulantes* using SMGC content. Most comparisons indicate a SMGC similarity of 40 percent and higher. Remarkably, clade III and clade IV share more SMGC families with species of clade I than between each other, suggesting that deletion events defined the SMGC content of these clades when they diverged from a common ancestor. Our results point to a general conservation of SMGC in section *Nidulantes*. Previous studies in section *Nigri* showed that species can have drastically different SMGC content which was as low as 10–20%. Here we find species to have at least 20–30% similarity while most comparisons are showing at 40–50%. This supports that clades III and IV are inside section *Nidulantes* which was still discussed (8, 47).

Secondary metabolites in section *Nidulantes*. Analytical studies of Aspergilli have shown a large chemical diversity through sections (48) with only a small fraction of compounds being conserved over large phylogenetic distances. In many cases SMs have been used to identify species (49). Hence, we were interested in the diversity of known secondary metabolite gene clusters throughout section *Nidulantes*. *Aspergillus* and *Penicillium* SMGC from the Minimum Information on Biosynthetic Gene clusters (MIBiG) database (50) were compared to SMGC of the dataset using protein BLAST. We identified best hits using a percent identity, query coverage and subject coverage cutoff of 95%. Since we predict members of a family to produce related compounds we annotated the SMGC family in a guilt by association approach — a process we term genetic dereplication.

Our analysis revealed several conserved clusters. The sterigmatocystin, fellutamide B (51), emericellamide, YWA, endocrocin (52), emericellin as well as the asperthecin SMGC (53) have related gene clusters in almost all species of section *Nidulantes*. Thus, they are characteristic for the section. The alternatiol, cichorine, F9775, asperfuranone, azanigerone and terrequinone cluster homologs are present in many clade I members and scattered over clade III and IV. This pattern indicates that the gene clusters have been acquired before differentiation of the *Nidulantes* section, then lost in some species during their development.

Apart from the conserved gene clusters in section *Nidulantes*, we identified SMGCs that are only present in a few species and suggest recent acquisition by horizontal gene transfer. A gene cluster in *A. multicolor* shows high synteny conservation to the hexadehydro-astechrome (HAS) gene cluster of *A. fumigatus*, whose product increases virulence (54), and to the astechrome gene cluster in *A. terreus* and 8. The latter cluster only produces astechrome since it misses a flavin adenine dinucleotide (FAD) binding enzyme (54, 55). The *A. multicolor* cluster, however, contains the FAD gene, suggesting it to produce HAS. To confirm the hypothesis of HGT, we identified a syntenic site to the area surrounding the HAS-homolog-locus of *A. multicolor* in *A. nidulans* (Fig. 8). We sustained our hypothesis of HGT comparing a ML phylogeny of HAS homologs against homologs of the conserved NRPS, SidC. Homologs of SidC in *A. multicolor*, *A. fumigatus* and *A. terreus* show a greater distance than their HAS homologs (Fig. 8). Furthermore, the distances shown for their HAS homologs resemble that of conserved NRPSs inside the same section, suggesting recent aquisition of the HAS cluster. Thus, we show evidence for a HGT of the HAS NRPS from *A. fumigatus* to *A. multicolor*.

A related SMGC to the penicillin gene cluster can be found in several species of clade I and III, as well as *P. chrysogenum* (Fig. 6). Furthermore, we find related gene clusters for 3,5-dimethylorsellinic acid based meroterpenoids terretonin (56), a mycotoxin, and austinol/dehydroaustinol (57), andrastin A, a promising antitumor agent (58), in one family. The gene cluster in *A. stella-maris* (supplemental file clusters) resembles the gene cluster structure of the calidodehydroaustin gene cluster (59). Six known examples of gene clusters are found throughout *Nidulantes*, while four are completely absent in clade II and other gene clusters investigated in *A. nidulans* have scattered presence patterns throughout species — as for the aspyridone hybrid gene cluster. In summary, we can identify different SMGC classes in section *Nidulantes* than in previous analyses of section *Nigri*. While section *Nigri* tends to contain more 6-methylsalicilic acid synthase derived clusters (Yanuthones, secalonic acid), section *Nidulantes* contains more orsellinic acid derived gene clusters (i.e. F9775, dehydroaustinol). The predicted SMGC for F9775 in *A. karnatakaensis* is producing a related compound which is further modified to a karnatakafuran (60), confirming our guilt by association approach. The detection of the sterigmatocystin gene clusters is in accordance with previous reports (14) proving the robustness of our method. Generally, SMGCs investigated in *A. nidulans* are found throughout section *Nidulantes*, emphasizing that they

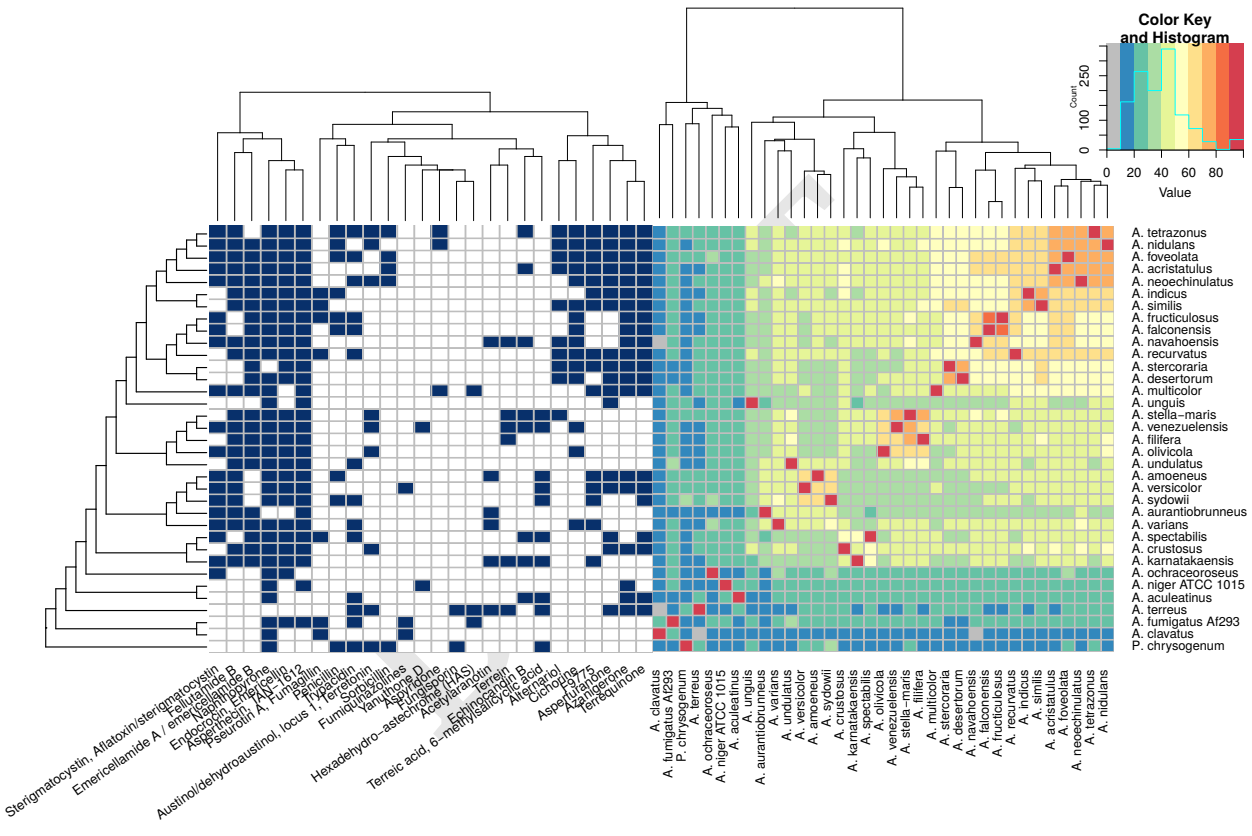


Fig. 6. SMGC diversity throughout the dataset. Left: Annotation of SMGC families with known compounds from the MIBiG database. Right: Comparison of SMGC families between species. The heatmap shows the amount of shared SMGC families in percent indicated by cell color. Column dendrogram of organisms is clustered hierarchically according to SMGC families.

are conserved. Some species show known SMGC from species of other sections, e.g. the hexadehydro-astechrome SMGC or the acetylaranotin. Using our comparative genomics approach we can identify horizontal gene transfers of gene clusters faster than on a single case basis (61, 62).

PKS and NRPS show diversity in closely related species. The differences in abundance of known gene clusters led us to investigate how many non-redundant SMGC are added to the dataset per genome sequenced. Taking *A. nidulans* SMGC families as the starting point, we identified the number of non-homologous SMGC families by addition of new species (Fig. 7). The plot is divided into clades as defined in figure 1. The first group shows the addition of new SMGC families from reference species — most new SMGC families are added here. Following this, gene clusters of species from section Nidulantes are added. The second group includes clades *A. aenei* and *A. versicolor*. Showing that organisms in the same section can still add to the non-redundant repertoire of all classes of SM genes. The third group consists of SMGC families from the clade II. PKS and NRPS families increase drastically. NPRS-Like, PKS-Like, HYBRID, TC, DMATS only increase slightly. The last group consists of clade I species. Only new NRPS, PKS and PKS-like are added to the set. Hence, only the most abundant classes show new families in closely related species. PKS and NRPS gene clusters seem to be the most abundant and diverse secondary metabolite genes on a section level. This indicates that these classes are subject to horizontal gene transfer or, in the case of NRPS, extensive domain rearrangement events.

Our genetic dereplication analysis showed that elucidated gene clusters of *A. nidulans* are biased towards section *Nidulantes* and clade I. Despite this, it is still informative to mine closely related species for new SMGC. As the analysis of non-redundant SMGC added per genome shows, we can expect that the species sequenced in this study will supply novel SMGC. Although most new SMGC are gained from Aspergilli of different sections, closely related species still yield a vast amount of non-redundant PKS and NRPS.

Conclusion

We *de novo* sequenced species in section *Nidulantes* and related their genetic content to the genome of the model organism *A. nidulans*. Our analysis shows that general regulators characterized in *A. nidulans* are distributed throughout its section. Protein family analysis indicated that in some cases we identified paralogs, pointing to a different function. In the case of LaeA paralogs this could mean that they function on different SMGCs than LaeA. Key cell cycle, cytoskeletal, and polarity genes are present throughout all species confirming *A. nidulans* role as model organism. Protein families provide insights into clade specific adaptations with loss of proteins in clade III and IV. Gene cluster predictions were confirmed by data on analytical studies in the case of sterigmatocystin and karnatakafurans. Hence guilt by association based annotation of general regulators and SMGCs is a reliable method to newly sequenced fungi.

Materials and Methods

Data collection. Protein, gff, smurf, interpro and go data was collected from JGI(A customized version of SMURF (63) was used to annotate secondary metabolite gene clusters throughout draft *Aspergillus* genomes. Protein sequences, smurf, interpro, GO and gff annotations were obtained from jgi (https://genome.jgi.doe.gov/).

Creating protein families.. Protein families were created using single linkage on bidirectional protein BLAST (64) hits with a percent identity of at least 50% and sum of coverage (query and subject) of at least 130% (27).

Whole genome phylogeny. Whole genome phylogeny was constructed from alignment of 200 bidirectional best hits between species using RAxML (23), MAFFT (24), and Gblocks (25).

Creating SMGC families.. SMGC families were created according to (27). In brief, bidirectional blast hits between proteins of secondary metabolite gene clusters were aggregated into a cluster vs cluster similarity score network. Subsequently, random walk clustering was used on this network to create SMGC families (65).

Genetic dereplication. MIBiG files were handled using biopython (66) and subsetted for entries from *Aspergillus* and *Penicillium* species. Selected entries were compared to secondary metabolite proteins of the dataset using protein BLAST (64) and best hits, which suffice a 95% pident, 95% query coverage and 95% subject coverage cutoff, annotated in our dataset.

ML phylogenies of NRPS. Protein sequences were aligned using clustalo (67), trimmed with trimal (68) using a gapthreshold of 0.8, a similarity threshold of 0.001, keeping 80% of columns. ML phylogenies were created using iqtree (69) with substitution model chosen according to ModelFinder Plus (70) and 1000 times ultrafast bootstrap (71).

1. Momany M (2002) Polarity in filamentous fungi: establishment, maintenance and new axes. *Current opinion in microbiology* 5(6):580–5.
2. Goldman GH, Kafer E (2004) Aspergillus nidulans as a model system to characterize the DNA damage response in eukaryotes. *Fungal Genetics and Biology* 41(4):428–442.
3. Virag A, Lee MP, Si H, Harris SD (2007) Regulation of hyphal morphogenesis by cdc42 and rac1 homologues in Aspergillus nidulans. *Molecular Microbiology* 66(6):1579–1596.
4. Shukla N, Osmani AH, Osmani SA (2017) Microtubules are reversibly depolymerized in response to changing gaseous microenvironments within Aspergillus nidulans biofilms. *Molecular Biology of the Cell* 28(5):634–644.
5. de Waard MA, van Nistelrooy JG (1979) Mechanism of resistance to fenarimol in Aspergillus nidulans. *Pesticide Biochemistry and Physiology* 10(2):219–229.
6. Osherov N, Kontoyiannis DP, Romans A, May GS (2001) Resistance to itraconazole in Aspergillus nidulans and Aspergillus fumigatus is conferred by extra copies of the A. nidulans P-450 14alpha-demethylase gene, pdmA. *The Journal of antimicrobial chemotherapy* 48(1):75–81.
7. Rocha EMF, Almeida CB, Martinez-Rossi NM (2002) Identification of genes involved in terbinafine resistance in Aspergillus nidulans. *Letters in Applied Microbiology* 35(3):228–232.
8. Hubka V, et al. (2016) A reappraisal of Aspergillus section Nidulantes with descriptions of two new sterigmatocystin-producing species. *Plant Systematics and Evolution* 302(9):1267–1299.
9. Engelhart S, et al. (2002) Occurrence of toxigenic Aspergillus versicolor isolates and sterigmatocystin in carpet dust from damp indoor environments. *Applied and Environmental Microbiology* 68(8):3886–3890.
10. Carmona EC, et al. (2005) Production, purification and characterization of a minor form of xylanase from Aspergillus versicolor. *Process Biochemistry* 40(1):359–364.
11. Alker AP, Smith GW, Kim K (2001) Characterization of Aspergillus sydowii (Thom et Church), a fungal pathogen of Caribbean sea fan corals. *Hydrobiologia* 460:105–111.
12. Terao K (1983) Sterigmatocystin-a masked potent carcinogenic mycotoxin. *Toxin Reviews* 2(1):77–110.
13. Veršilovskis A, de Saeger S (2010) Sterigmatocystin: Occurrence in foodstuffs and analytical methods - an overview. *Molecular Nutrition and Food Research* 54(1):136–147.
14. Rank C, et al. (2011) Distribution of sterigmatocystin in filamentous fungi. *Fungal Biology* 115(4-5):406–420.
15. Raper KB (1946) the Development of Improved Penicillin-Producing Molds. *Annals of the New York Academy of Sciences* 48(2):41–56.
16. Holt G, MacDonald KD (1968) Isolation of strains with increased penicillin yield after hybridization in Aspergillus nidulans. *Nature* 219(5154):636–637.
17. Diez B, Ii V, Martin JF, Barredosll JL (1990) The Cluster of Penicillin Biosynthetic Genes. *Biochemistry* 265(27):16358–16365.
18. Nielsen ML, et al. (2011) A genome-wide polyketide synthase deletion library uncovers novel genetic links to polyketides and meroterpenoids in Aspergillus nidulans. *FEMS Microbiol Lett* 321(2):157–166.
19. Klejnstrup ML, et al. (2012) *Genetics of Polyketide Metabolism in Aspergillus nidulans.* Vol. 2, pp. 100–133.

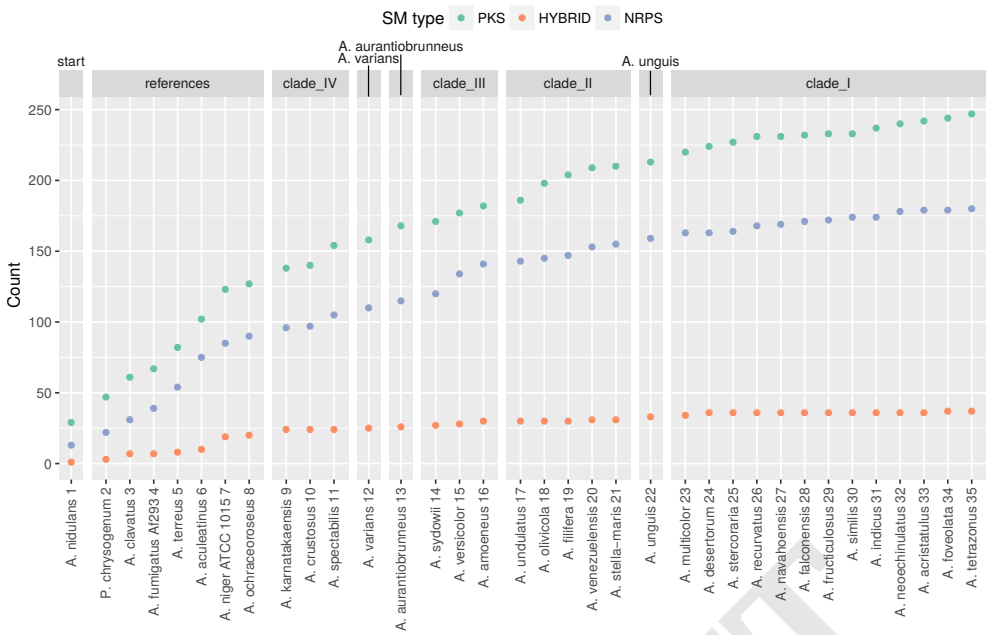


Fig. 7. Non-redundant SMGC added per genome. Scatterplot showing the number of non-redundant SMGC added per genome; starting with *A. nidulans* SMGC families. The type of SMGC is color coded. Blue rectangles indicate groups of (from left to right) reference species, *A. aenei* and *A. versicolor* clade, *A. stellatus* clade, and *A. nidulans* clade.

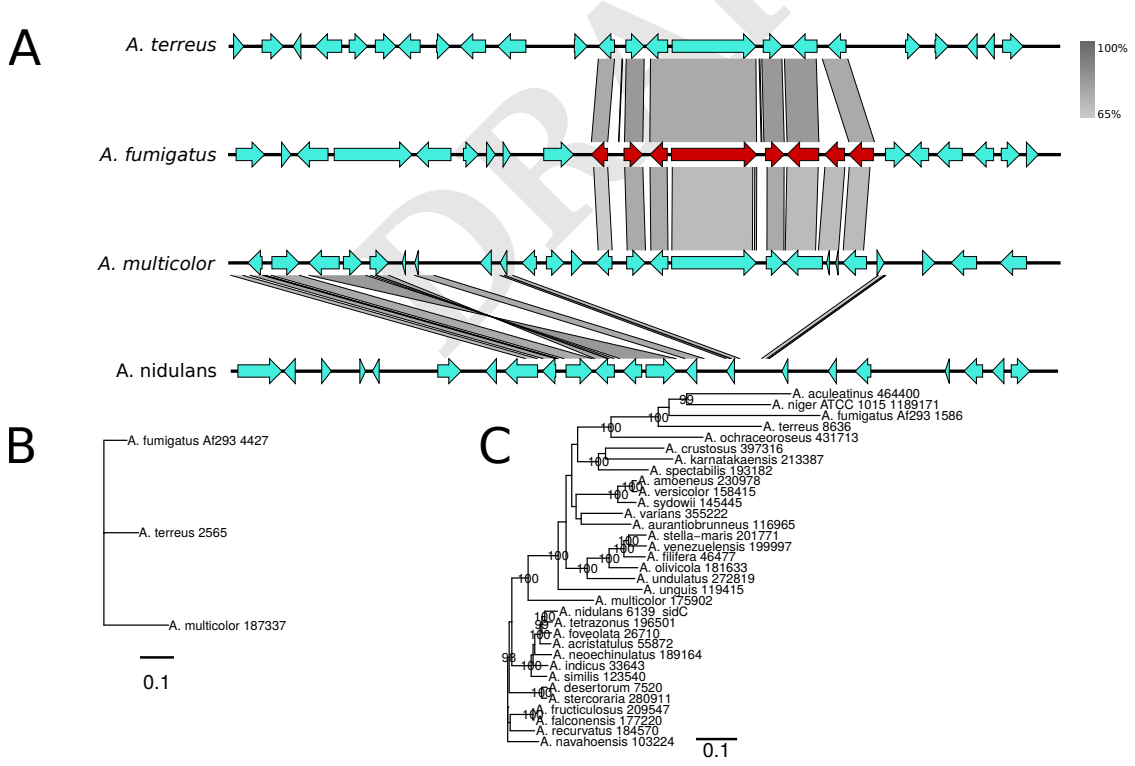


Fig. 8. synteny between Aspergilli. **A** Synteny plot for HAS locus in *A. terreus*, *A. fumigatus* and *A. multicolor*, and best hit locus in *A. nidulans*. hasA-H (54) are conserved in *A. multicolor*, while hasG is missing in A. terreus. Thus we expect *A. multicolor* to synthesize full HAS (*A. terreus* only produces astechrome according to Bok et al. (55)). *A. nidulans* contains a region which is syntenic to the upstream region of the HAS locus. Hence, we expect that the HGT event occured at Chr VII of *A. nidulans*. Homologs for these HAS proteins were only found in the family. **B** ML phylogeny of HAS and (**C**) sidC-homolog NRPSs. Protein sequences were aligned using clustal omega and trimmed using trimal prior to maximum likelihodd analysis using iqtree. HAS homologs (left) show phylogenetic proximity. Comparison to distances of a conserved

20. Sanchez JF, et al. (2010) Molecular genetic analysis of the orsellinic acid/F9775 genecluster of Aspergillus nidulans. *Mol. BioSyst.* 6(3):587–593.
21. Galagan JE, et al. (2005) Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. *Nature* 438(7071):1105–1115.
22. Chen A, et al. (2016) Aspergillus section Nidulantes (formerly Emericella): Polyphasic taxonomy, chemistry and biology. *Studies in Mycology* pp. 1–118.
23. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
24. Katoh K, Misawa K, Kuma Ki, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* 30(14):3059–3066.
25. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution* 17(4):540–552.
26. de Vries RP, et al. (2016) *Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus Aspergillus.* pp. 1–45.
27. Vesth TC, et al. (2018) The genomes of Aspergillus section Nigri reveal drivers in fungal speciation (in preparation). *In preparation.*
28. Ries LNA, Beattie SR, Espeso EA, Cramer RA, Goldman GH (2016) Diverse Regulation of the CreA Carbon Catabolite Repressor in Aspergillus nidulans. *Genetics* 203(1):335–352.
29. Denison SH (2000) pH regulation of gene expression in fungi. *Fungal Genetics and Biology* 29(2):61–71.
30. Bok JW, Keller NP (2004) LaeA , a Regulator of Secondary Metabolism in Aspergillus spp. *Eukaryotic Cell* 3(2):527–535.
31. Shroff RA, Lockington RA, Kelly JM (1996) Analysis of mutations in the creA gene involved in carbon catabolite repression in Aspergillus nidulans. *Can J Microbiol* 42(9):950–959.
32. Oakley CE, et al. (2016) Discovery of McrA, a master regulator of Aspergillus secondary metabolism. *Molecular microbiology* 103(November 2016):347–365.
33. Christensen U, et al. (2011) Unique regulatory mechanism for D-galactose utilization in Aspergillus nidulans. *Applied and Environmental Microbiology* 77(19):7084–7087.
34. Futagami T, et al. (2011) Putative stress sensors WscA and WscB are involved in hypo-osmotic and acidic pH stress tolerance in aspergillus nidulans. *Eukaryotic Cell* 10(11):1504–1515.
35. Oakley CE, et al. (2017) Discovery of McrA, a master regulator of Aspergillus secondary metabolism. *Molecular Microbiology* 103(2):347–365.
36. Harris SD, Momany M (2004) Polarity in filamentous fungi: moving beyond the yeast paradigm. *Fungal genetics and biology : FG & B* 41(4):391–400.
37. Riquelme M (2013) Tip Growth in Filamentous Fungi: A Road Trip to the Apex. *Annual Review of Microbiology* 67(1):587–609.
38. Si H, Rittenour WR, Harris SD (2016) Roles of Aspergillus nidulans Cdc42/Rho GTPase regulators in hyphal morphogenesis and development. *Mycologia* 108(3):543–555.
39. Harris SD, Hamer L, Sharpless KE, Hamer JE (1997) The Aspergillus nidulans sepA gene encodes an FH1/2 protein involved in cytokinesis and the maintenance of cellular polarity. *The EMBO journal* 16(12):3474–83.
40. Doshi P, Bossie CA, Doonan JH, May GS, Morris NR (1991) Two alpha-tubulin genes of Aspergillus nidulans encode divergent proteins. *Molecular & general genetics : MGG* 225(1):129–41.
41. Kirk KE, Morris NR (1991) The tubB alpha-tubulin gene is essential for sexual development in Aspergillus nidulans. *Genes & development* 5(11):2014–23.
42. Pan F, Malmberg RL, Momany M (2007) Analysis of septins across kingdoms reveals orthology and new motifs. *BMC evolutionary biology* 7:103.
43. Hernández-Rodríguez Y, et al. (2014) Distinct septin heteropolymers co-exist during multicellular development in the filamentous fungus Aspergillus nidulans. *PloS one* 9(3):e92819.
44. Nishihama R, Onishi M, Pringle JR (2011) New insights into the phylogenetic distribution and evolutionary origins of the septins. *Biological Chemistry* 392(8-9):681–7.
45. De Souza CP, et al. (2013) Functional Analysis of the Aspergillus nidulans Kinome. *PLoS ONE* 8(3).
46. de Vries RP, Visser J, Ronald P, de Vries, R., P. (2001) Aspergillus Enzymes Involved in Degradation of Plant Cell Wall Polysaccharides. *Microbiology and Molecular Biology Reviews* 65(4):497–522.
47. Kocsubé S, et al. (2016) Aspergillus is monophyletic: Evidence from multiple gene phylogenies and extrolites profiles. *Studies in Mycology* 85:199–213.
48. Frisvad JC, Larsen TO (2015) Chemodiversity in the genus Aspergillus. *Applied Microbiology and Biotechnology* 99(19):7859–7877.
49. Frisvad JC, Andersen B, Thrane U (2008) The use of secondary metabolite profiling in chemotaxonomy of filamentous fungi. *Mycological Research* 112(2):231–240.
50. Medema MH, et al. (2015) Minimum Information about a Biosynthetic Gene cluster. *Nature Chemical Biology* 11(9):625–631.
51. Yeh Hh, et al. (2016) Resistance Gene-Guided Genome Mining: Serial Promoter Exchanges in Aspergillus nidulans Reveal the Biosynthetic Pathway for Fellutamide B, a Proteasome Inhibitor.
52. Lim FY, et al. (2012) Genome-based cluster deletion reveals an endocrocin biosynthetic pathway in Aspergillus fumigatus. *Applied and Environmental Microbiology* 78(12):4117–4125.
53. Szewczyk E, et al. (2008) Identification and characterization of the asperthecin gene cluster of Aspergillus nidulans. *Applied and Environmental Microbiology* 74(24):7607–7612.
54. Yin WB, et al. (2013) A nonribosomal peptide synthetase-derived iron(III) complex from the pathogenic fungus aspergillus fumigatus. *Journal of the American Chemical Society* 135(6):2064–2067.
55. Bok JW, et al. (2015) Fungal artificial chromosomes for mining of the fungal secondary metabolome. *BMC Genomics* 16(1):1–10.
56. Guo CJ, et al. (2012) Molecular genetic characterization of a cluster in A. terreus for biosynthesis of the meroterpenoid terretonin. *Organic Letters* 14(22):5684–5687.
57. Lo HC, et al. (2012) Two separate gene clusters encode the biosynthetic pathway for the meroterpenoids austinol and dehydroaustinol in Aspergillus nidulans. *Journal of the American Chemical Society* 134(10):4709–4720.
58. Matsuda Y, Awakawa T, Abe I (2013) Reconstituted biosynthesis of fungal meroterpenoid andrastin A. *Tetrahedron* 69(38):8199–8204.
59. Valiante V, et al. (2017) Discovery of an Extended Austinoid Biosynthetic Pathway in Aspergillus calidoustus. *ACS Chemical Biology* 12(5):1227–1234.
60. Manniche S, Sprogøe K, Dalsgaard PW, Christophersen C, Larsen TO (2004) Karnatakafurans A and B: Two dibenzofurans isolated from the fungus aspergillus karnatakaensis. *Journal of Natural Products* 67(12):2111–2112.
61. Khaldi N, Collemare J, Lebrun MH, Wolfe KH (2008) Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. *Genome biology* 9(1):R18.
62. Lawrence DP, Kroken S, Pryor BM, Arnold AE (2011) Interkingdom gene transfer of a hybrid NPS/PKS from bacteria to filamentous Ascomycota. *PloS one* 6(11):e28231.
63. Khaldi N, et al. (2010) SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal genetics and biology : FG & B* 47(9):736–41.
64. Camacho C, et al. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10(1):421.
65. Pons P, Latapy M (2005) Computing communities in large networks using random walks. *Physics and Society* p. arXiv:physics/0512106.
66. Cock PJ, et al. (2009) Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11):1422–1423.
67. Sievers F, et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology* 7(1):539.
68. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972–1973.
69. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32(1):268–274.
70. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS (2017) ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates. *Nature Methods* 14(6):587–591.
71. Minh BQ, Nguyen MAT, Von Haeseler A (2013) Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and Evolution* 30(5):1188–1195.

Chapter 6

Conclusions and perspectives

Individual comparisons on gene to gene basis were feasible because there were only few genome sequences available. Now projects like the *Aspergillus* sequencing project or the 1K fungal genome project provide a vast amount of information — creating a need for new methods. Comparative genomics and secondary metabolite gene networks are capable of processing this information and provide new insights into fungal evolution and biology.

This thesis highlighted how comparative genomics can be used to identify gene dynamics over a large set of organisms and how it can be used to characterize *de novo* sequenced species. Chapter 2 presented how comparative genomics can be used to describe species on the section, clade and isolate level and determine the drivers for fungal speciation.

Chapter 3 showed how secondary metabolite gene cluster (SMGC) networks can be used to characterize SMGCs through many species at once. The method points out related SMGCs for analogous compounds in different species — a valuable tool to identify new leads for bioactive compounds. This also highlights how SMGCs evolved and how *Aspergilli* create new SMs. Finally, combining production data of SMs with gene cluster families led to the identification of MlfA, the NRPS producing the anit-cancer enhancing compound malformin. This is an important step to support the elucidation of SMGCs for desired compounds. Aspergilli have been investigated for their SMs during many years, offering a vast amount of data on produced compounds. Creating SMGC families of newly sequenced Aspergilli and combining them with the available metabolite data will provide many more SMGC leads for pharmaceutically relevant compounds. Chapter 5 prove that the established pipeline is applicable to other sections and scalable to more genomes. Here, we also show how protein families can be used to characterize newly sequenced species with annotations available from model organisms. The general regulators of secondary metabolism and their paralogs, which have been identified in our analysis, can serve as leads for overexpression

studies.

The thesis also described how genus wide protein sets can be used to characterize the evolutionary history of SM proteins. Our analysis of PKS-NRPS and NRPS-PKS proteins pointed towards multiple events leading to their evolution — a new insight in hybrid phylogenies. The phylogeny which was established in chapter 4 will serve as a lead for new studies to reengineer hybrids. The field of combinatorial biosynthesis, which tries to create chimeras of these enzymes, has been hampered by the lack of leads of hybrids amenable to recombination.

Generally, this thesis (and the manuscripts included in it) will serve as a go-to reference for future work on sections *Nigri* and *Nidulantes*. Characterizing their proteome in protein families enabled us to provide an overview of the distribution of each protein of interest throughout all species. Thus, instead of individual searches by BLAST others might use our dataset and the analyses in order to find a protein of interest. Providing the distribution of proteins throughout all species might provide others with key insights they need. Furthermore, the method and content on diversity and dynamics of SMGCs are transferable to other genera and will enable studies on analogous SM pathways in fungi. Imitating SMGC dynamics seen for analogous compounds might pave the way for synthetic biology approaches.

Appendix A

Appendix

A.1 Supplementary information: The genomes of *Aspergillus* section *Nigri* reveal drivers in fungal speciation

Materials and methods

Fungal strains

Unless otherwise noted, the species examined were taken from the IBT Culture Collection of Fungi at the Technical University of Denmark (DTU). Strains employed in this study are denoted in SI 1.

Purification of DNA and RNA

For all sequences generated for this study (SI 1), spores were defrosted from storage at -80°C and inoculated onto solid CYA medium. Fresh spores were harvested after 7-10 days and suspended in a 0.1% Tween solution. Spores were stored in solution at 5°C for up to three weeks. For generation of biomass, spores were inoculated in sterile liquid CYA medium and cultivated for 5-10 days at 30°C. Fungal mycelium was isolated from the liquid medium by filtration through Miracloth and flash-frozen in liquid nitrogen. DNA isolation was performed using a modified version of the standard phenol extraction [Green, Michael R., and Joseph. Sambrook. Molecular Cloning : Cold Spring Harbor Laboratory Press, 2012] and checked for quality and concentration using a NanoDrop (BioNordika, DK). RNA isolation was performed using the Qiagen RNAEasy Plant Mini Kit according to the manufacturer's instructions.

Biomass for all fungal strains was obtained from shake flasks containing 200 mL of complex media CYA, meaox, or CY20 depending on the strain (see SI 1). Biomass was isolated by filtering through Miracloth (Millipore, 475855-1R), freeze dried, and stored at -80°C. A sample of frozen biomass was subsequently used for RNA purification. First, hyphae were transferred to a 2 mL microtube, together with a 5mm steel bead (QIAGEN), placed in liquid nitrogen, then lysed using the QIAGEN TissueLyser LT at 45 Hz for 50 seconds. Then the QIAGEN RNeasy mini Plus Kit was used to isolate RNA., RLT Plus buffer (with 2-mercaptoethanol) was added to the samples, vortexed, and spun down. The lysate was then used in step 4 in the instructions provided by the manufacturer, and the protocol was followed from this step. For genomic DNA, a protocol based on Fulton et al. [Fulton, TM. “Microprep protocol for extraction of DNA from tomato and other herbaceous plants.” Plant Molecular Biology Reporter 13.3 (1995): 207–209] was used.

Protocol for preparation of fungal DNA

The protocol described below has successfully been employed to isolate genome-sequencing grade genomic DNA for more than 200 different *Aspergillus* species.

List of Materials

- D-Sorbitol (Sigma, S1876 – CAS 50-70-4)
- Tris-Base (Sigma 7-9, T1378 – CAS 7786-1)
- 37% HCl (Th. Geyer, 836,1000)
- EDTA (Merck, 324503 – CAS 6381-92-6)

- Sodium chloride (NaCl) (AppliChem A1371,9010 – CAS 7647-14-5]
- Cetyl trimethylammonium bromide (CTAB) (Sigma 52365 – CAS 57-09-0)
- Sarkosyl NL (Sigma, L5777 – CAS 137-16-6)
- Polyvinylpyrrolidone (PVP) (Sigma PVP-40T – CAS 9003-39-8)
- Proteinase K (NEB P8107S)
- Potassium acetate (J. T. Baker 0129910025 – CAS 127-08-2)
- Phenol:Chloroform:Isoamylalcohol (25:24:1) (Sigma P3803)
- Sodium acetate (J. T. Baker 9914011001 – CAS 6131-90-4]
- 96 % Ethanol (vwr chemicals)
- 70 % Ethanol (vwr chemicals)
- Isopropanol (Merck, 109634 – CAS 67-63-0)
- Liquid nitrogen
- Sodium hydroxide (Sigma S5881 – CAS 1310-73-2)
- RNase A (Sigma R-4875 – CAS 9001-99-4)

Preparation of liquid media

All solutions were autoclaved.

Buffers

- 5M Potassium acetate (pH 7.5)
    - 122.5 g potassium acetate and ddH_2O up to 250 mL
    - pH adjusted with acetic acid
- 3M Sodium acetate
    - 81.65 g sodium acetate
    - Add ddH_2O up to 200 mL
- 1% PVP
    - 2 g PVP in 200 mL ddH_2O
- 5% Sarkosyl
    - 10 g Sarkosyl in 200 mL ddH_2O
- 1M Tris-HCl (pH 9)
    - 60.57 g Tris-base and 4.81 mL 37% HCl
    - Add ddH_2O up to 500 mL
- 0.5M EDTA
    - 116.4 g EDTA
    - Add ddH_2O up to 500 mL
    - Add sodium hydroxide pellets until pH reaches 8.0
- Buffer A
    - 31.9 g sorbitol, 50 mL 1M Tris-HCl (pH 9), 5 mL 0.5M EDTA (pH 8)
    - Add ddH_2O up to 500 mL
- Buffer B
    - 100 mL 1M Tris-HCl (pH 9), 50 mL 0.5M EDTA, 58.44 g NaCl, 10 g CTAB
    - Add ddH_2O up to 500 mL
- TE (pH 9)
    - 1.21 g Tris-base, 0.37 g EDTA
    - Add ddH_2O up to 1000 mL
- Lysis Buffer (for 10 mL pr. Sample)
    - 3.75 mL Buffer A
    - 3.75 mL Buffer B

    - 1.5 mL 5 % Sarkosyl
    - 1 mL 1 % PVP
    - 100 µl Proteinase K
- RNase A
    - Dissolve 10 mg dry powder in 1 mL ddH_2O

Equipment

- NanoDrop ND 1000 Spectrophotometer or NanoDrop Lite From Qiagen
- Qubit 1.0 fluorometer from Invitrogen and Qubit dsDNA BR Assay Kit (Q32853) from ThermoFisher.
- Mortar and pestle
- Centrifuge for 50 mL Falcon tubes at 4°C

Protocol

1. Pre-heat Buffer B to 65°C.
2. Prepare Lysis Buffer just before use and keep at 65°C.
3. Transfer freeze-dried mycelia into a mortar and cover with liquid nitrogen.
    a. Grind material and transfer to a 50 mL Falcon tube as soon as all liquid nitrogen has evaporated.
    b. Powder in the tube should not exceed the 5 mL mark, but a minimum of 3 mL is recommended.
    c. Note powder must not thaw.
4. Add 10 mL Lysis Buffer and mix vigorously by vortexing.
5. Incubate for 30 minutes at 65°C.
    a. Mix frequently by inverting the tube.
6. Add 3.35 mL 5 M potassium acetate.
    a. Mix gently by inverting the tube 5-7 times.
    b. Incubate solution 30 minutes on ice.
7. Centrifuge for 30 minutes at 5,000 g at 4°C.
8. Transfer the supernatant (~ 9 mL) to a new 50 mL Falcon tube and add 5 mL of phenol:chloroform:isoamylalcohol (25:24:1)
    a. Mix gently 5-7 times.
9. Centrifuge 20 minutes at 4,000 g at 4 °C.
10. Transfer the aqueous phase (~8 mL) to a new 50 mL Falcon tube.
    a. Note avoid any material from the interphase.
11. Add 100 µl RNase A (10 mg/mL) and mix gently.
    a. Incubate at room temperature for 30-60 minutes
12. Add 1/10 volume of 3M sodium acetate and 1 volume ice-cold 96% ethanol. (Alternatively, isopropanol can be used, but it may adversely influence A260/A280 measurements).
13. Incubate solution at 20°C for 30 minutes.
14. Centrifuge for 30 minutes at 10,000 g and 4°C.
15. Discard the supernatant.
16. Wash the pellet with 2 mL 70 % ethanol and pipette as much away without disturbing the pellet.
17. Dry the pellet at room temperature until all ethanol has evaporated (approximately 15 minutes).

18. Note: do not let the pellet dry out!
19. Dissolve the pellet in 500 µL TE. This may take an over-night incubation at room temperature with light shaking. Transfer DNA solution to a 2 mL Eppendorf tube.
20. Take an aliquot for DNA quality assessments (see below) and store the remaining DNA solution at -20 °C until further use.
21. For testing DNA quality:
22. Make a 20-fold dilution of the DNA solution (from step 19) in a 1.5 mL Eppendorf tube to a total volume of 100 µl.
23. Run 5-10 µl of the diluted sample on an agarose gel to estimate the quality and concentration.
24. Use the nanodrop for A_{260}/A_{280} measurements. Ratios should be in the range of 1.6-2.2.
25. Use the Qubit to determine DNA concentration estimations. Good solutions fall in the range of 20-200 ng/µl DNA in stock solution.

DNA and RNA sequencing and assembly

All genomes in this study, except for *A. heteromorphus, A. eucalypticola, and A. sclerotioniger*, and all transcriptomes were sequenced with Illumina. The genomes of *A. heteromorphus, A. eucalypticola, and A. sclerotioniger* were sequenced with PacBio.

For all genomic Illumina libraries, 100 ng of DNA was sheared to 270 bp fragments using the Covaris LE220 (Covaris) and size selected using SPRI beads (Beckman Coulter). The fragments were treated with end-repair, A- tailing, and ligated to Illumina compatible adapters (IDT, Inc) using the KAPA-Illumina library creation kit (KAPA biosystems).

For transcriptomes, stranded cDNA libraries were generated using the Illumina Truseq Stranded RNA LT kits. mRNA was purified from 1 µg of total RNA using magnetic beads containing poly-T oligos. mRNA was fragmented using divalent cations and high temperature. The fragmented RNA was reverse transcribed using random hexamers and SSII (Invitrogen) followed by second strand synthesis. The fragmented cDNA was treated with end-pair, A-tailing, adapter ligation, and 10 cycles of PCR.

The prepared libraries were quantified using KAPA Biosystem's next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. The quantified libraries were then multiplexed with other libraries, and library pools were prepared for sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq paired-end cluster kit, v3, and Illumina's cBot instrument to generate clustered flowcells for sequencing. Sequencing of the flowcells was performed on the Illumina HiSeq2000 sequencer using a TruSeq SBS sequencing kit, v3, following a 2x150 indexed run recipe.

After sequencing, the genomic fastq files were QC filtered to remove artifacts/process contamination and assembled using Velvet (Zerbino DR, Birney E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18(5):821-829.). The resulting assemblies were used to create *in silico* long mate-pair libraries with inserts of 3000 +/-

90 bp, that were then assembled with the target fastq using AllPathsLG release version R47710, (Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A. 25;108(4).). Illumina transcriptome reads were assembled into consensus sequences using Rnnotator v. 3.3.2 (Martin J, Bruno VM, Fang Z, Meng X, Blow M, Zhang T, Sherlock G, Snyder M, Wang Z. (2010) Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genomics 2010 11(1), 663.).

For the genomes of *A. heteromorphus*, *A. eucalypticola*, and *A. sclerotioniger,* amplified libraries were generated using a modified shearing version of Pacific Biosciences standard template preparation protocol. 5 µg of gDNA was used to generate each library. The DNA was sheared using a Covaris LE220 focused-ultrasonicator with their Red miniTUBES to generate fragments of 5 kb in length. The sheared DNA fragments were then prepared according to the Pacific Biosciences protocol using their SMRTbell template preparation kit, where the fragments were treated with DNA damage repair (ends repaired so that they were blunt-ended and 5' phosphorylated). Pacific Biosciences hairpin adapters were then ligated to the fragments to create the SMRTbell template for sequencing. The SMRTbell templates were then purified using exonuclease treatments and size-selected using AMPure PB beads.

Sequencing primer was then annealed to the SMRTbell templates and Version P4 sequencing polymerase was bound to them. The prepared SMRTbell template libraries were sequenced on a Pacific Biosciences RSII sequencer using Version C2 chemistry and 2 hour sequencing movie run times. The three PacBio genomes datasets were assembled using HGAP3 (http://files.pacb.com/software/smrtanalysis/2.2.0/doc/smrtportal/help/!SSL!/Webhelp/CS_Prot_RS_HGAP_Assembly3.htm).

All genomes were annotated using the JGI annotation Pipeline (Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, Smirnova T, Nordberg H, Dubchak I, Shabalov I. (2014) MycoCosm portal: gearing up for 1000 fungal genomes. Nucleic Acids Res. 42(1):D699-704.).
Genome assembly and annotations are available at the JGI fungal genome portal MycoCosm (Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, Smirnova T, Nordberg H, Dubchak I, Shabalov I. (2014) MycoCosm portal: gearing up for 1000 fungal genomes. Nucleic Acids Res. 42(1):D699-704.) (http://jgi.doe.gov/fungi) and have been deposited at DDBJ/EMBL/GenBank under the following accessions (TO BE PROVIDED UPON PUBLICATION).

Cultivation for secondary metabolite analysis

Fungal strains were cultivated as 3-point cultures on Czapek yeast extract agar (CYA), Czapek yeast extract agar with 5% NaCl (CYAS) [Nielsen, Kristian Fog et al. "Review of Secondary Metabolites and Mycotoxins from the *Aspergillus niger* Group." Analytical and Bioanalytical

Chemistry 395.5 (2009): 1225–1242], yeast extract sucrose agar (YES) media for 7 days in the dark at 25˚C. Three 6 mm ID plugs taken across of the cultures were then extracted using an (3:2:1) (ethyl acetate/dichloromethane/methanol) mixture and dissolved in methanol (Smedsgaard, J. “Micro-Scale Extraction Procedure for Standardized Screening of Fungal Metabolite Production in Cultures.” Journal of Chromatography a 760.2 (1997): 264–270.).

Chemical analysis of secondary metabolites

All chemical analyses were done by reversed phase ultra high performance liquid chromatography (UHPLC) coupled to UV-Vis diode array detection (DAD) combined with either fluorescence detection (FLD) or high resolution mass spectrometry (HRMS). Three different methods were used:

Method 1

UHPLC-DAD Pure UHPLC-DAD-FLD was performed using a Dionex RSLC Ultimate 3000 (Dionex, Sunnyvale, CA) system linked to an Agilent 1100 FLD detector (Agilent Technologies, Santa Clara, CA). The system was equipped with an Agilent Poroshell Phenyl-hexyl column (150 × 2.1mm, 2.6 mm), and run using a linear gradient of water–acetonitrile starting from 10% to 100% (both containing 50 ppm trifluoroacetic acid) over 8 minutes, then 100% acetonitrile for 2 minutes. The column temperature was 60˚C, the flow rate 0.8 mL/min, and the injection volume was 1 µL. The UV spectra 200-640 nm were matched against our internal database [Nielsen, Kristian Fog et al. “Review of Secondary Metabolites and Mycotoxins from the *Aspergillus niger* Group.” Analytical and Bioanalytical Chemistry 395.5 (2009): 1225–1242].

Method 2

UHPLC-DAD-HRMS was conducted on a Dionex RSLC Ultimate system linked to maXis HD QTOF MS (Bruker Daltonics, Bremen Germany). Separation was done on a Kinetex C18 column (100 × 2.1mm, 2.6 mm), with a linear gradient consisting of water and acetonitrile (both buffered with 20 mM formic acid), starting at 10% acetonitrile and increased to 100% in 10 minutes where it was held for 2 minutes, returned (0.4 mL/min, 40°C). Injection volume, depending on sample concentration, typically varied between 0.1-1 µL. Samples were analyzed in ESI+ and some also in ESI- full scan mode scanning m/z 100-1250. Data were analyzed by aggressive dereplication [Klitgaard, Andreas et al. “Aggressive Dereplication Using UHPLC–DAD–QTOF: Screening Extracts for up to 3000 Fungal Secondary Metabolites.” Analytical and Bioanalytical Chemistry 406.7 (2014): 1933–1943] using lists of compounds considered to be from black Aspergillus only (~350), a lists with all Aspergillus compounds (~2450), as well as a list of 1600 reference standards, of which 500 are known to come from Aspergillus. Unknown peaks were matched against Antibase2012 and dereplicated using accurate mass, isotope patterns, adduct patterns, logD, and UV/Vis data (Klitgaard, A. “Aggressive Dereplication Using UHPLC-DAD-QTOF: Screening Extracts for up to 3000 Fungal Secondary Metabolites.” (2016)).

Method 3

UHPLC-DAD-HRMS was conducted on an Agilent Infinity 1290 UHPLC system coupled to an Agilent 6550 QTOF MS. Separation was obtained on an Agilent Poroshell 120 phenyl-hexyl

column (2.1 x 250 mm, 2.7 µm) using a linear gradient of water and acetonitrile (both buffered with 20 mM formic acid), from 10% to 100% acetonitrile in 15 minutes, where it was held for 2 min. The flow was 0.35 mL/min and temperature 60°C. Injection volume was between 0.1-1 µL depending on the sample concentration.

Samples were analyzed in ESI+ and some also in ESI-full scan mode scanning m/z 100-1700 and with automatic MS/MS enabled for ion counts above 100 000 counts and with a quarantine time of 0.06 minutes. MS/MS spectra were obtained at 10, 20 and 40 eV [Kildgaard, Sara et al. “Article Accurate Dereplication of Bioactive Secondary Metabolites from Marine-Derived Fungi by UHPLC-DAD-QTOFMS and a MS/HRMS Library.” (2015)].

Full scan data were analyzed as above in Masshunter [Kildgaard, Sara et al. “Accurate Dereplication of Bioactive Secondary Metabolites from Marine-Derived Fungi by UHPLC-DAD-QTOFMS and a MS/HRMS Library.” Marine Drugs 12.6 (2014): 3681–3705]. MS/MS data were matched to our internal MS library (~1700 compounds) of reference standards and tentatively identified compounds [Kildgaard, Sara et al. “Accurate Dereplication of Bioactive Secondary Metabolites from Marine-Derived Fungi by UHPLC-DAD-QTOFMS and a MS/HRMS Library.” Marine Drugs 12.6 (2014): 3681–3705].

Genome annotations

Genome annotation

All genomes were annotated based on the JGI annotation pipeline [Grigoriev, Igor V., Diego A. Martinez, and Asaf A. Salamov. “Fungal Genomic Annotation.” Applied Mycology and Biotechnology 6.C (2006): 123–142] as previously described [Kis-Papo, Tamar et al. “Genomic Adaptations of the Halophilic Dead Sea Filamentous Fungus Eurotium Rubrum.” Nature Communications 5 (2014): 3745].

Whole-genome phylogeny

Protein sequences of all organisms were compared using BLASTp (e-value cutoff 1e-05). Orthologous groups of sequences were constructed based on best bidirectional hits (BBH). Two hundred groups with a member from each species were selected and the sequences of each organism was concatenated into one long protein sequence. Concatenated sequences are aligned using MAFFT (thread 16) and well-aligned regions were extracted using Gblocks (-t = p -b4=5 -b5 = h). Trees were then constructed using multithreaded RAxML, the PROTGAMMAWAG model and 100 bootstrap replicates.

Prediction of Secondary Metabolite Gene Clusters

For the prediction of secondary metabolite (SM) clusters, we developed a command-line Python script roughly following the SMURF algorithm [Khaldi, N. et al. SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet. Biol. 47, 736–41 (2010)]. As input, the program takes genomic coordinates and the annotated PFAM domains of the predicted genes.

Based on the multi-domain PFAM composition of identified 'backbone' genes, it can predict seven types of SM clusters: 1) Polyketide synthases (PKSs), 2) PKS-like, 3) nonribosomal peptide-synthetases (NRPSs) 4) NRPS-like, 5) hybrid PKS-NRPS, 6) prenyltransferases (DMATS), and 7) terpene cyclases (TCs). Besides backbone genes, PFAM domains, which are enriched in experimentally identified SM clusters (SM-specific PFAMs) were used in determining the borders of gene clusters. The maximum allowed size of intergenic regions in a cluster was set to 3 kb and each predicted cluster was allowed to have up to 6 genes without SM-specific domains.

Prediction of Secreted Proteases

Secretome prediction was done using an in-house adaptation of SignalP [Bendtsen, JD et al. “Improved Prediction of Signal Peptides: SignalP 3.0.” Journal of Molecular Biology 340.4 (2004): 783–795].

Identification of Fungal Metabolites

Fungal metabolite extracts were prepared using one of the three following methods (Nielsen, Kristian Fog et al. “Review of Secondary Metabolites and Mycotoxins from the *Aspergillus niger* Group.” Analytical and Bioanalytical Chemistry 395.5 (2009): 1225–1242):

- Chloroform-methanol-acetone-ethylacetate extraction
- Micro-extraction using methanol-dichloromethane-ethyl acetate
- 75% methanol

Metabolites were analyzed by LC-DAD, LC-DAD-TOFMS, or LC-MS. Data was analyzed based on LC retention times, UV spectra or MS data and compared to an in-house adaptation of Antibase [Klitgaard, Andreas et al. “Aggressive Dereplication Using UHPLC–DAD–QTOF: Screening Extracts for up to 3000 Fungal Secondary Metabolites.” Analytical and Bioanalytical Chemistry 406.7 (2014): 1933–1943].

Gene-Compound Assignment

Identification of conserved or highly similar fungal gene clusters was performed based on the gene cluster predictions above. The genomes were compared using the BLASTp function from the BLAST+ suite [Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).]. Presence/absence of an orthologous gene to a member in a gene cluster was based on a bi-directional best hit, with e<1e-100 and coverage of >90%. Presence/absence of a full gene cluster was based on the occurrence of the majority of the predicted members in a gene cluster, including the backbone synthetase in another species.

Detection of encoded CAZymes

Each Aspergillus protein model was compared using BLASTp to proteins listed in the Carbohydrate-Active Enzymes database (www.cazy.org, [Lombard, Vincent et al. “The Carbohydrate-Active Enzymes Database (CAZy) in 2013.” Nucleic Acids Research 42.1 (2014): D490–D495]). Models with over 50% identity over the entire length of an entry in CAZy were directly assigned to the same family (or subfamily when relevant). Proteins with less than 50%

identity to a protein in CAZy were all manually inspected and conserved features, such as the catalytic residues, were searched whenever known. Because an identical 30% sequence identity results in widely different e-values (from non-significant to highly significant), for CAZy family assignments, we examine sequence conservation (percentage identity over CAZy domain length). Sequence alignments with isolated functional domains were performed in the case of multimodular CAZymes. The same methods were used for *Penicillium chrysogenum* and *Neurospora crassa*..

Genetic diversity

Mapping of genes shared by groups of species

All predicted sets of protein sequences for the 38 genomes analyzed were aligned using the BLASTp function from the BLAST+ suite version 2.2.27 (e-value <= 10^-10 cutoff). These 1.444 whole-genome BLAST tables were analyzed to identify bidirectional hits in all pairwise comparisons. Using custom Python-scripts, homologs were identified within and across the genomes and grouped into sequence similar families using single linkage, if they met? the following criterion: \The sum of the alignment coverage between the pairwise sequences ≥ 130%, the alignment identity between the pairwise sequences ≥ 50%, and the hit must be found in both of the species BLAST output (reciprocal hits). Singletons were assigned a family having only one gene member. This allowed for identification of species unique genes, genes shared by sections, clades and sub-clades of species. All homologs were assigned functional and structural domains using InterPro version 48 [Jones, Philip et al. “InterProScan 5: Genome-Scale Protein Function Classification.” Bioinformatics 30.9 (2014): 1236–1240.] and checked for annotation and sequencing errors by investigating scaffold location and sequence identity.

For the analysis of the pan and core genomes of a subset of 38 fungal species used in this study, the orthologous and paralogous families were subsetted to include only the species of interest. Therefore the genes representing the core and unique portion of the genomes will adjust relative to the accompanying species.

Secondary metabolism

Our implementation of SMURF [Khaldi, N. et al. SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet. Biol. 47, 736–41 (2010)] was run on genomic data from 37 *Aspergillus* strains. Proteins of the resulting secondary metabolism gene clusters (SMGCs) were compared to each other by alignment using BLASTp (BLAST+ suite version 2.2.27, e-value <= 10^-10). Subsequently, a score based on BLASTp identity and shared proteins was created to determine the similarity between gene clusters as depicted in the formula below. Using these scores, we created a weighted network of SMGC clusters and used a random walk community detection algorithm (R version 3.3.2, igraph_1.0.1, [Pons, Pascal, and Matthieu Latapy. “Computing Communities in Large Networks Using Random Walks .” (2012)], cutoff 1 step) to determine families of SMGC clusters. Finally, we ran another round of random walk clustering on the communities which contained more members than species in the analysis.

$$\frac{ptailoring * ntailoring}{ttailoring} * 0.35 + \frac{pbackbone * nbackbone}{tbackbone} * 0.65$$

$$\frac{\sum (pident_{tailoring})}{nx_{tailoring}} * 0.35 + \frac{\sum (pident_{backbones})}{nmax_{backbones}} * 0.65$$

To create a cluster similarity score, a combined score of tailoring and backbone enzymes was created. The sum of the BLASTp percent identity (pident) of all hits for tailoring enzymes between two clusters was divided by the maximum amount of tailoring enzymes (nmax) and multiplied by 0.35. Then the score for the backbone enzymes was calculated in the same manner but multiplied by 0.65 to give more weight to the backbone enzymes. The scores were added to create an overall cluster similarity score.

$$avg(pident_{tailoring}) * 0.35 + avg(pident_{backbones}) * 0.65$$

Identification of shared SMGC families at nodes of the phylogenetic tree

A list containing organisms of each branch of the phylogenetic tree was created and compared to the list of organisms for each SMGC family. If all organisms of a family matched, the count on the corresponding node was increased by one.

Prediction of the Aurasperone B gene cluster

Lists of organisms for all SMGC families were compared to the lists of aurasperone B producing species and filtered for Interpro annotations containing the terms “cytochrome P450” or “methyltransferase”.

Primary metabolism

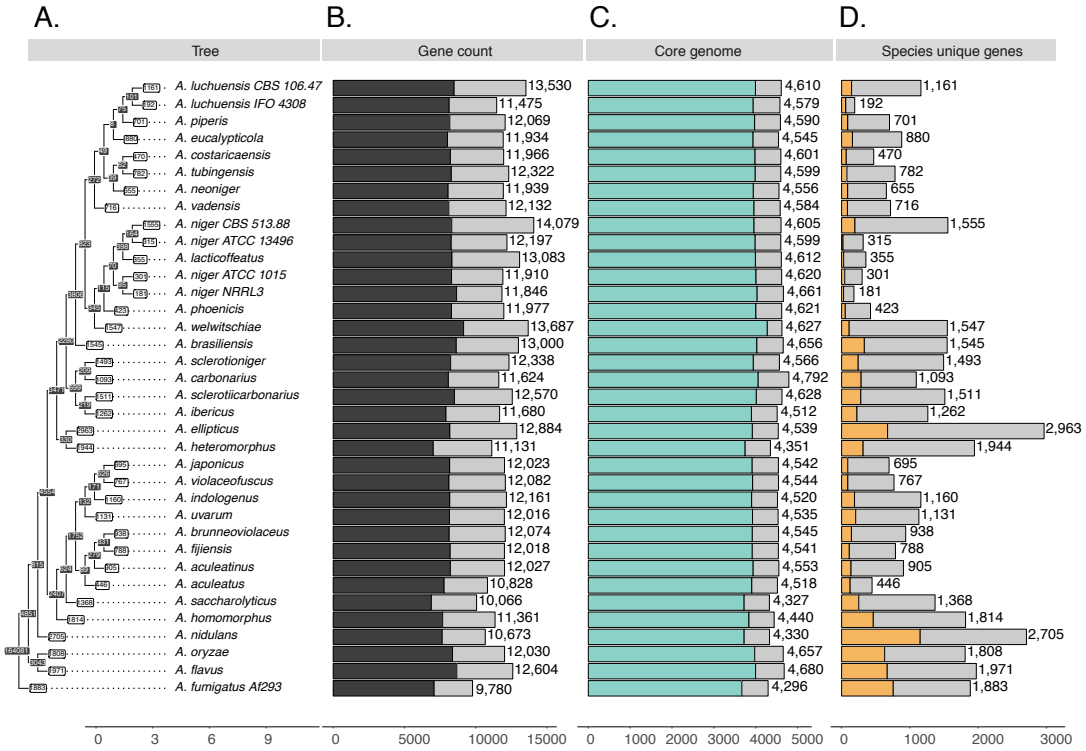
Copy numbers were assessed using the homologous protein families generated during the analysis of genome diversity. The gene pathway associations were taken from the *A. niger* genome-scale model [Andersen, Mikael Rørdam, Michael Lynge Nielsen, and Jens Nielsen. “Metabolic Model Integration of the Bibliome, Genome, Metabolome and Reactome of *Aspergillus niger*.” Molecular Systems Biology 4 (2008): 1–13]. All proteins in the respective protein families were considered putative isozymes and were included in the copy number analyses.

Comparing the putative isoenzymes in the different species, gene sequences were aligned and clustered using neighbor-joining with MUSCLE v3.8.31 [Edgar, Robert C. “MUSCLE: Multiple Sequence Alignment with Improved Accuracy and Speed.” Proceedings - 2004 Ieee Computational Systems Bioinformatics Conference, Csb 2004 (2004): 728–729]. Resulting trees have been visualized and edited for publication using the python ETE Toolkit [Huerta-Cepas, Jaime, Joaquin Dopazo, and Toni Gabaldon. “ETE: a Python Environment for Tree Exploration.” Bmc Bioinformatics 11.6 (2010): 24]. Subcellular localizations for the genes included in the

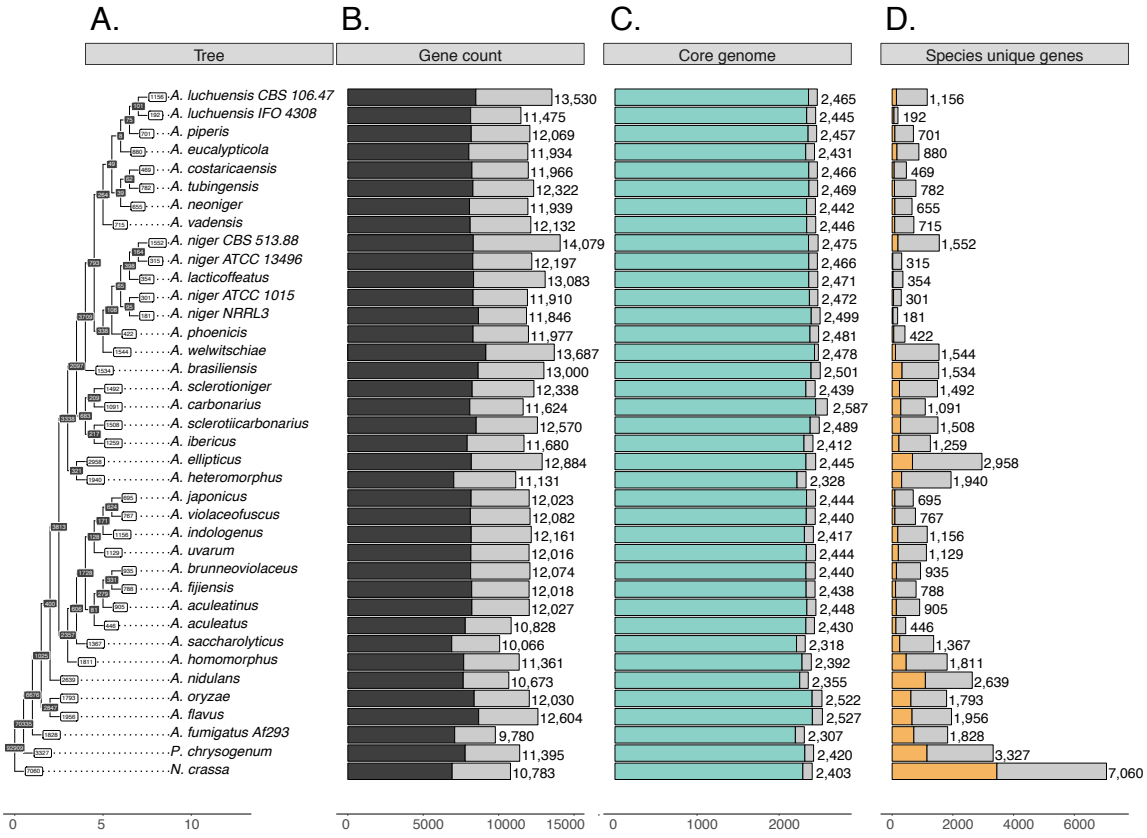
analysis were predicted using the TargetP [Emanuelsson, O et al. “Predicting Subcellular Localization of Proteins Based on Their N-Terminal Amino Acid Sequence.” Journal of Molecular Biology 300.4 (2000): 1005–1016] webserver at http://www.cbs.dtu.dk/services/TargetP/.

SI 3: Phylogenetic relation and InterPro coverage of the *Aspergillus* genus species. A. Dendrogram of the phylogenetic relation between the 36 species.

The black boxes represent the homologous genes among the species branching from the nodes. The white boxes represent the genes unique to the specific species. B. Genome sizes of each species. The black coloured boxes represent the genes annotated by InterPro. C. Core genome size of each species. The blue coloured boxes represent the genes annotated by InterPro. D Species unique genes. The yellow coloured boxes represent the genes annotated by InterPro. The grey coloured boxes represent genes with no annotation. The numbers at right side of the boxes indicates the total number of annotated and not annotated genes. The bar scales are unique to each graph.



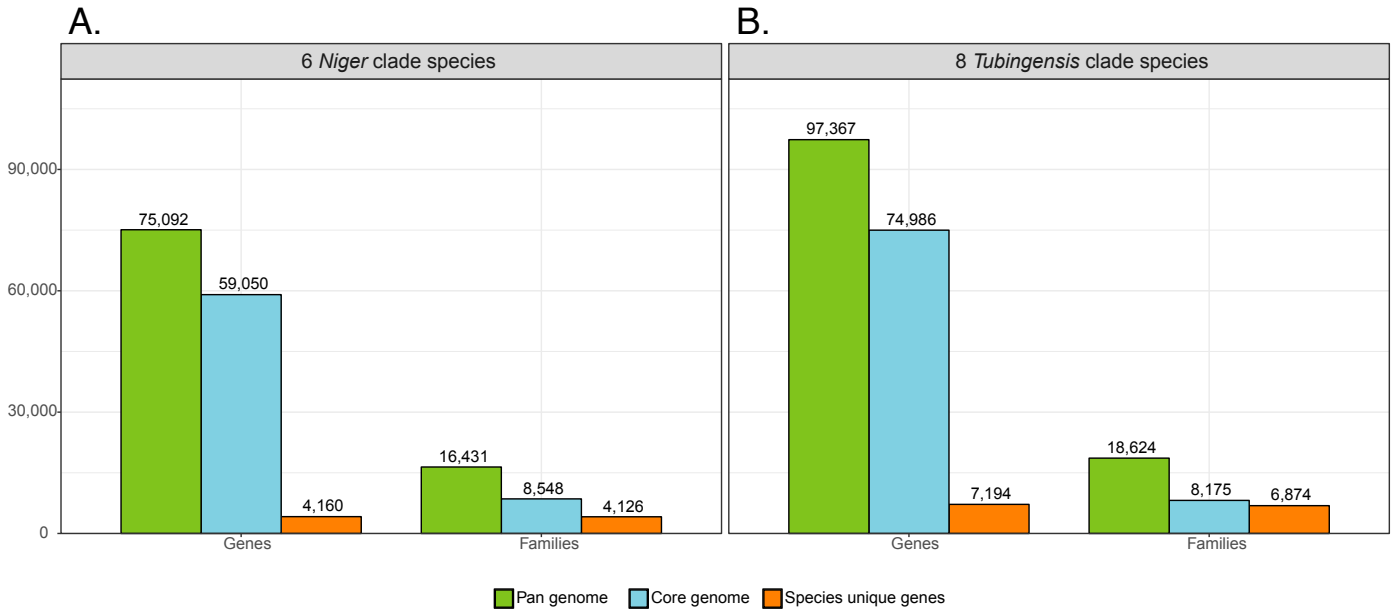
SI 5: Phylogenetic relation and InterPro coverage of the fungal species included in this study. A. Dendrogram of the phylogenetic relation between the 38 species. The black boxes represent the homologous genes among the species branching from the nodes. The white boxes represent the genes unique to the specific species. B. Genome sizes of each species. The black coloured boxes represent the genes annotated by InterPro. C. Core genome size of each species. The blue coloured boxes represent the genes annotated by InterPro. D Species unique genes. The yellow coloured boxes represent the genes annotated by InterPro. The grey coloured boxes represent genes with no annotation. The numbers at right side of the boxes indicates the total number of annotated and not annotated genes. The bar scales are unique to each graph.



SI 10

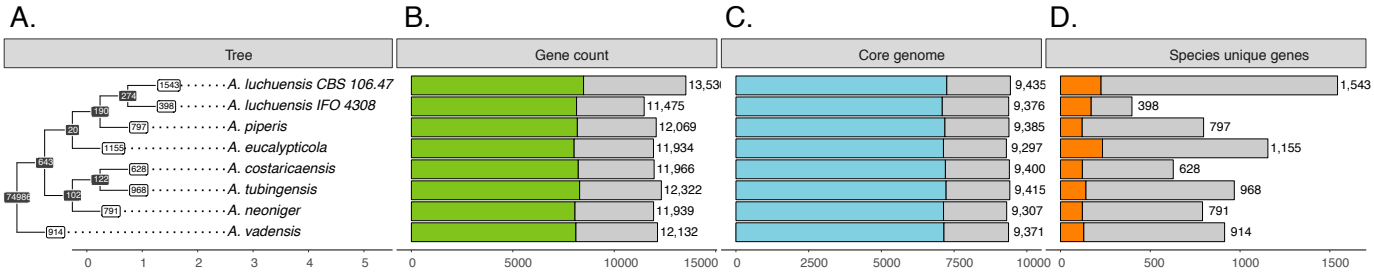
A. Pan, core, unique genes and families of the *Nigri* clade species
The total number of the proteins and families of all genes in the *Tubingensis* clade (pan genome, green), genes shared by all species (core genome, blue), and genes unique to the individual species (species unique genes, orange).The families were predicted using BLASTp alignments with cutoffs specific to the *Aspergillus* genus and single linkage clustering designed for this project. *Niger* clade species:*A. lacticoffeatus, A. niger ATCC 1015, A. niger ATCC 13496, A. niger CBS 513.88, A. niger NRRL3* and *A. phoenicis*.

B. Pan, core, unique genes and families of the Tubingensis clade species
The total number of the proteins and families of all genes in the Tubingensis clade (pan genome, green), genes shared by all species (core genome, blue), and genes unique to the individual species (species unique genes, orange).The families were predicted using BLASTp alignments with cutoffs specific to the *Aspergillus* genus and single linkage clustering designed for this project. *Tubingensis* clade species: *A. costaricaensis, A. eucalypticola, A. luchuensis CBS 106.47, A. luchuensis IFO 4308, A. neoniger, A. piperis, A. tubingensis* and *A. vadensis.*



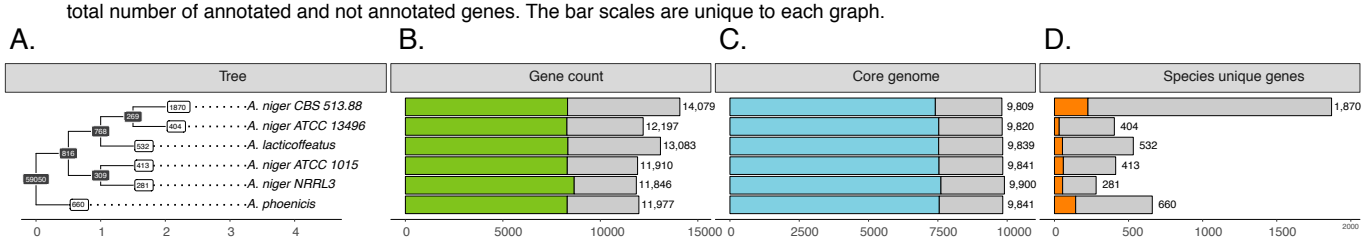
S11

Phylogenetic relation and InterPro coverage of the Tubingensis clade species. A. Dendrogram of the phylogenetic relation between the 8 species. The black boxes represent the homologous genes among the species branching from the nodes. The white boxes represent the genes unique to the specific species. B. Gene count of each species. The green coloured boxes represent the genes annotated by InterPro. C. Core genome size of each species. The blue coloured boxes represent the genes annotated by InterPro. D. Species unique genes. The orange coloured boxes represent the genes annotated by InterPro. The grey coloured boxes represent genes with no annotation. The numbers at right side of the boxes indicates the total number of annotated and not annotated genes. The bar scales are unique to each graph.



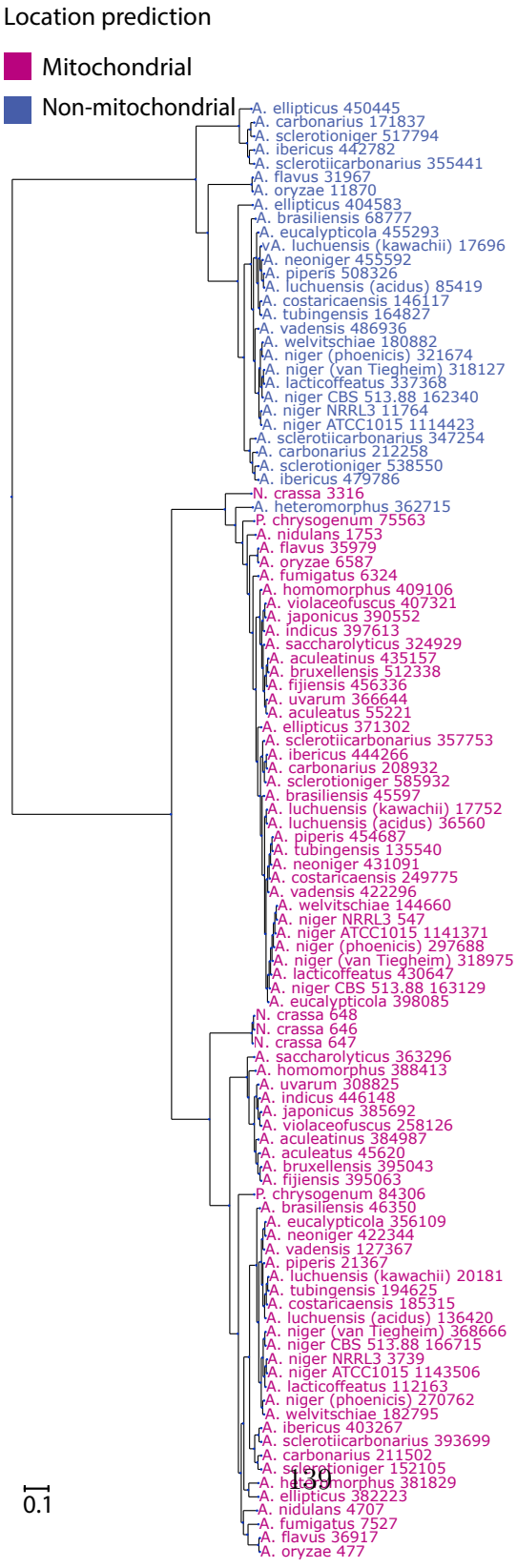
Dendrogram and InterPro coverage of the Niger clade species

Phylogenetic relation and InterPro coverage of the Niger clade species. A. Dendrogram of the phylogenetic relation between the 6 species. The black boxes represent the homologous genes among the species branching from the nodes. The white boxes represent the genes unique to the specific species. B. Gene counts of each species. The green coloured boxes represent the genes annotated by InterPro. C. Core genome size of each species. The blue coloured boxes represent the genes annotated by InterPro. D. Species unique genes. The orange coloured boxes represent the genes annotated by InterPro. The grey coloured boxes represent genes with no annotation. The numbers at right side of the boxes indicates the total number of annotated and not annotated genes. The bar scales are unique to each graph.

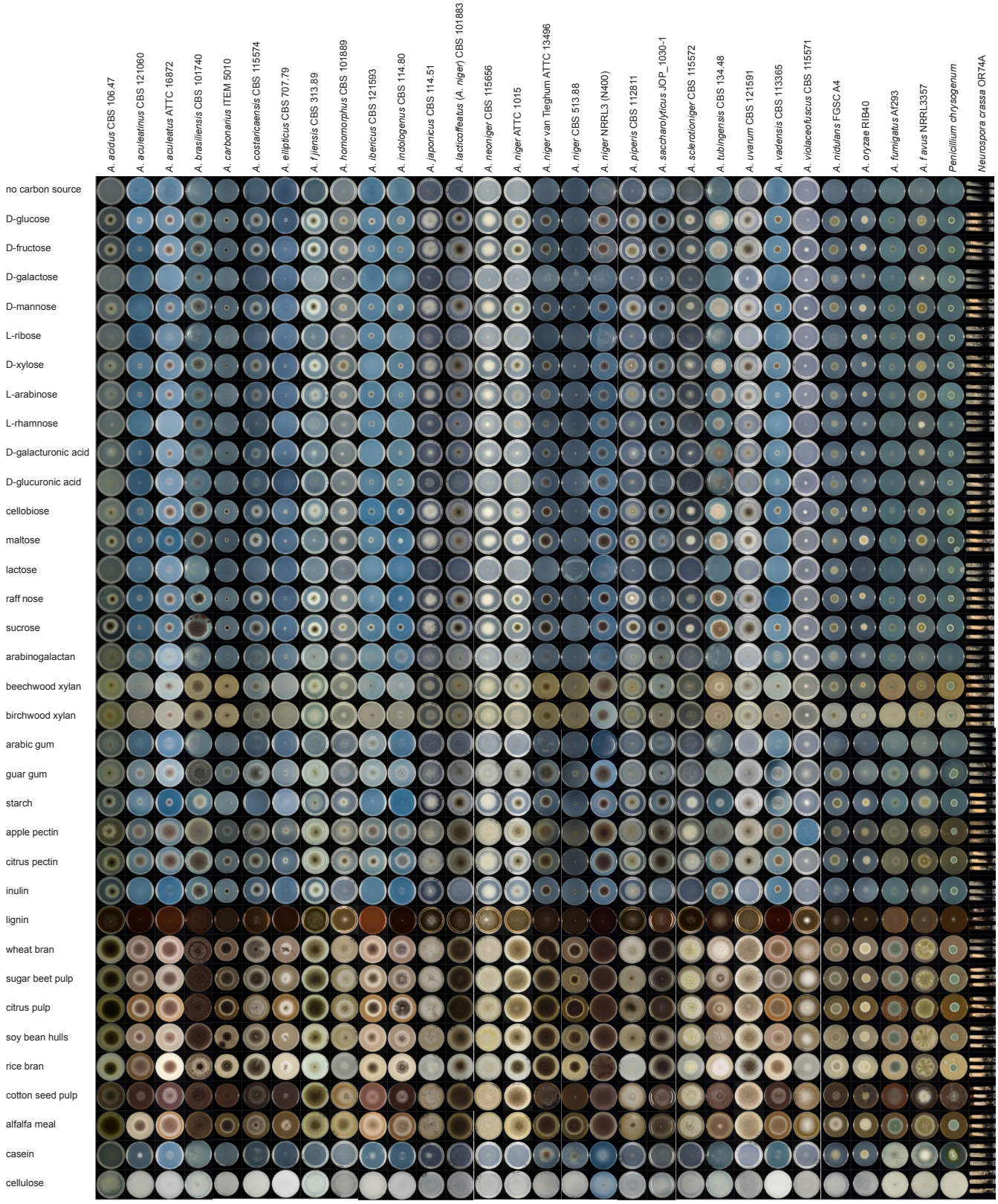


SI 14: Sequence comparison of citrate synthase genes

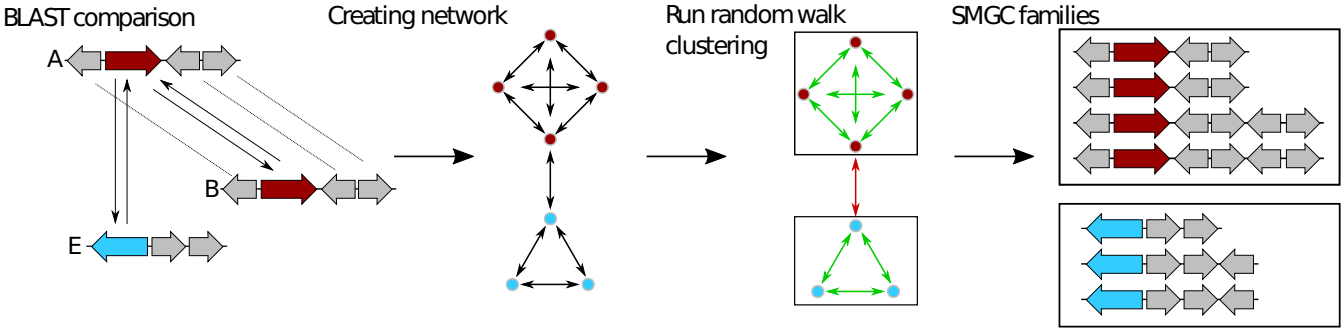
The sequence of the putative citrate synthases identified in the copy number analysis has been compared using neighbor-joining as implemented in MUSCLE. Potential subcellular locations have been predicted using the TargetP method.



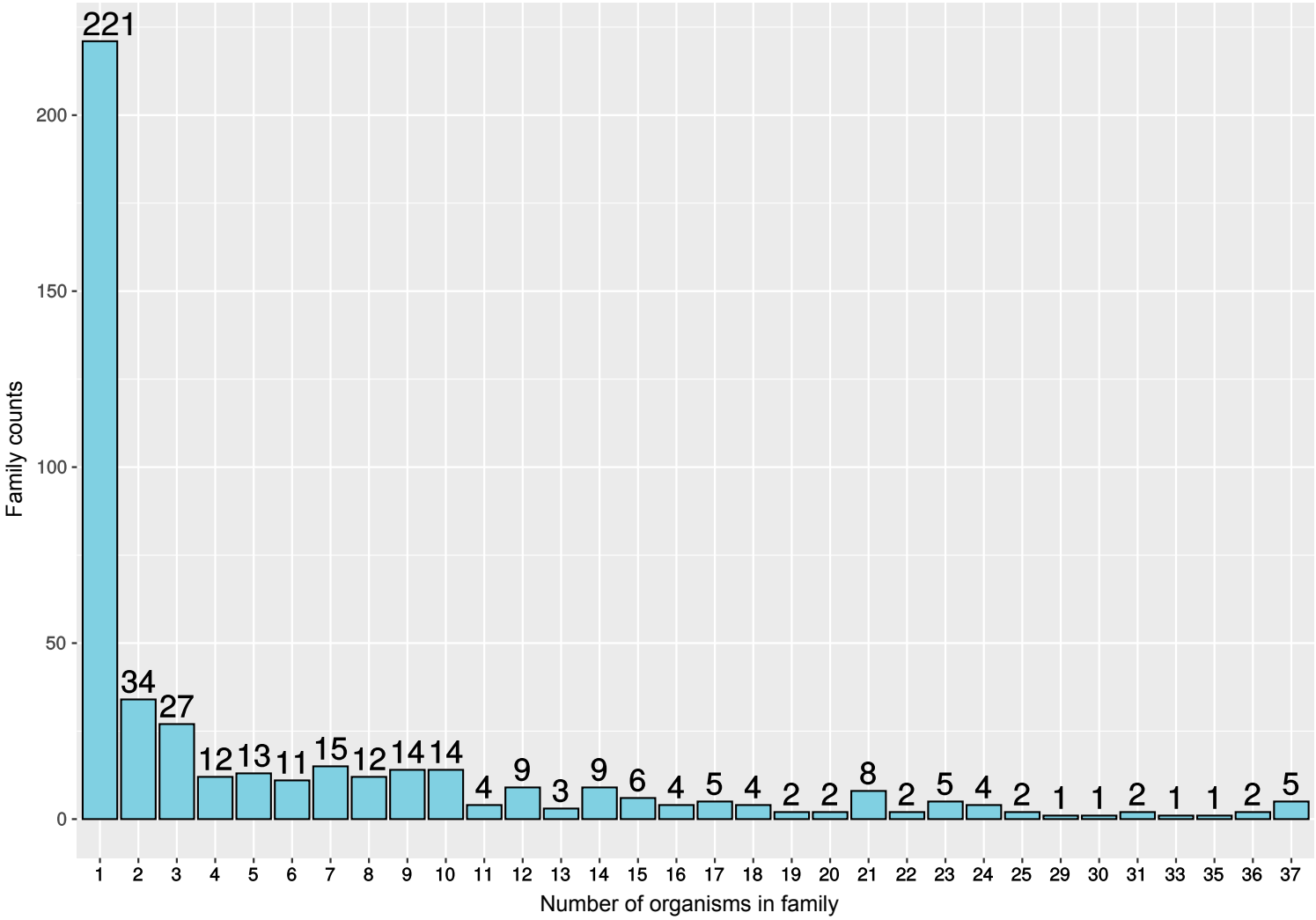
SI 17: Comparative growth profiling of Aspergilli from section Nigri and reference fungi.



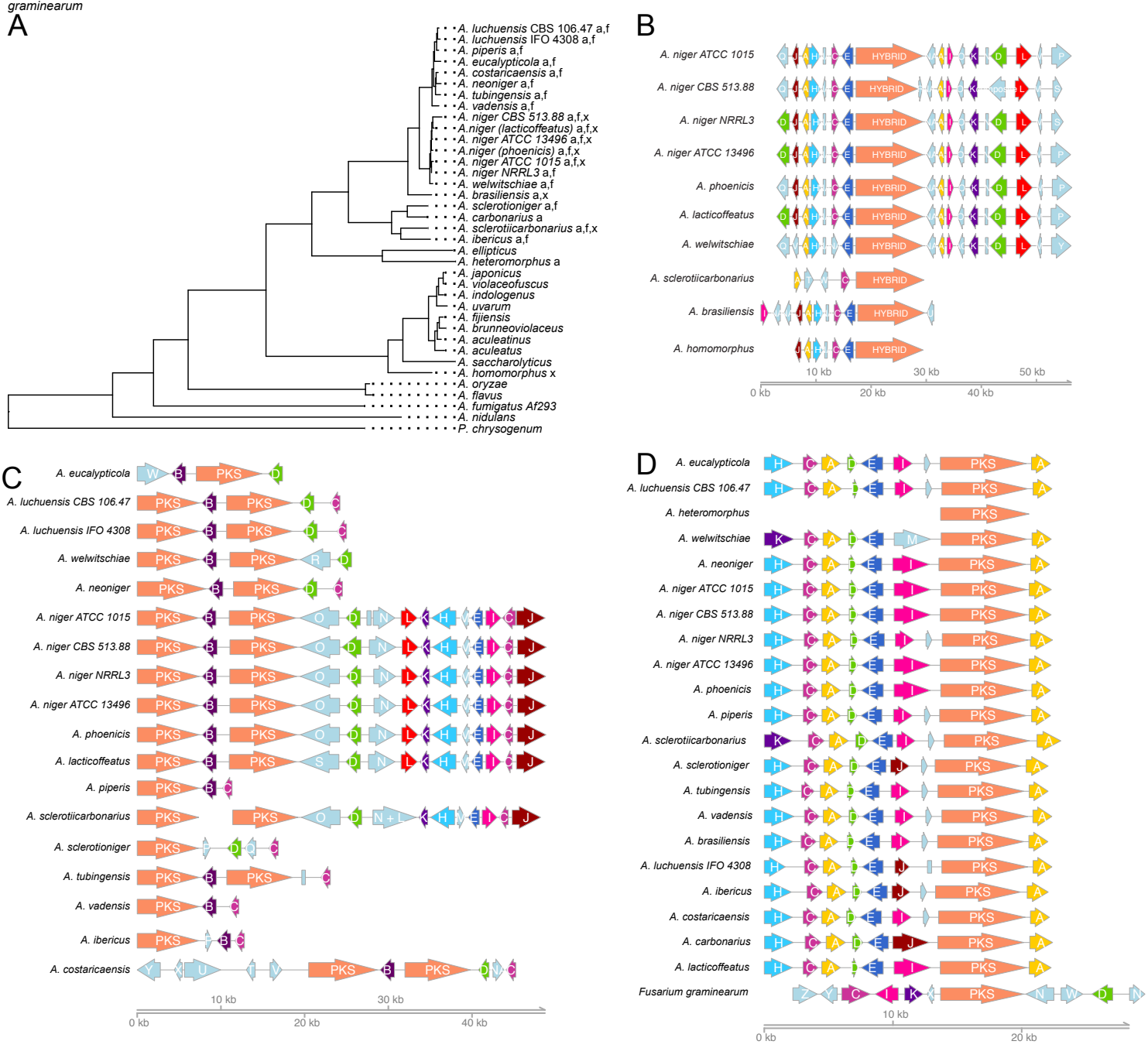
SI 18: Schematic representation of secondary metabolic gene cluster family identification. Protein blast comparisons between all gene cluster members are aggregated into one cluster similarity score. From this, a network is created with gene clusters as nodes (dots) and similarity score as edges (arrows). Subsequently, random walk clustering is used to find communities of nodes inside the network. Green arrows indicate a high probability that nodes will be assigned to one community. Red arrows indicate community borders. Resulting families are containing communities of related SMGC.



SI 19: Barplot describing SMGC family frequencies. The bars illustrate the presence of a SMGC family in a certain number of organisms. Numbers above bars show the total counts.



SI 20: Overview of predicted SMGC families (B-D) and their location in the phylogeny. A: Cladogram of species used for secondary metabolic gene cluster analysis in this study. Letter code indicates predicted clusters for fumonisin (f), aurasperone B (a) and an example gene cluster family (x) predicted exclusive ly in biseriates. B: Example gene cluster found in five distantly related species (x). C: Predicted gene cluster family (f) containing a PKS in *Aspergillus niger* CBS 513.8 similar to FUM 1 from *Fusarium oxysporum*. D: Predicted gene cluster family for aurasperone B (a) including the aurofusarin gene cluster from *Fusarium graminearum*



SI 21: Overview of SM families detected in 27 different *Aspergillus* isolates

Species	**Compound families**																																								
	Notoamides	Oxalines	Okaramins	Aurantiamins	Atromentins	Pyranonigrins	Tensidols	Fumonisins	Malformins	Kotanins	Ochratoxins	Geodins	Calbistrins	Aculenes	Aflavinins	Naphtho-gamma-pyrone	Yanuthones	Funalenones	Xanthoascins	Terphenyllins	Austdiols	Candidusins	Paspalinins	Asperazines	Secalonic acids	Emodins	Nigragillins	Corymbiferan lactones	Aspernomine	Asperflavins	Neopyranonigrins	Pyrophens	Dehydrocarolic acid	Sclerotionigerins	Decaturins
A. brunneoviolaceus	X	X	X	X																																				
A. lacticoffeatus					X	X	X	X	X	X	X																														
A. uvarum			X									X	X	X	X																										
A. niger NRRL3						X	X	X	X	X						X	X	X																							
A. ellipticus					X														X	X	X	X																		
A. saccharolyticus														X									X																	
A. tubingensis						X			X							X		X						X																
A. aculeatus													X	X											X	X														
A. piperis						X									X	X																								
A. niger CBS 513.88						X	X	X	X		X					X		X									X													
A. neoniger					X	X									X	X		X									X	X	X											
A. violaceofuscus				X									X																	X	X								
A. brasiliensis					X		X		X							X																X	X						
A. eucalypticola					X	X										X		X																					
A. luchuensis CBS 106.47						X										X		X									X												
A. sclerotiicarbonarius						X										X							X																
A. costaricaensis															X	X		X										X	X										
A. japonicus				X																										X								
A. heteromorphus					X								X	X																									
A. aculeatinus		X		X									X											X	X				X	X									
A. ibericus					X	X									X	X											X												
A. sclerotioniger					X						X					X		X									X									X			
A. vadensis										X						X		X						X			X												
A. fijiensis	X	X		X									X												X	X				X	X								
A. homomorphus																									X					X			X		X			
A. niger ATCC 1015						X	X	X	X	X						X		X									X												
A. indologenus			X												X																								

A.2 Supplementary information: Uncovering bioactive compounds in *Aspergillus* section *Nigri* by genetic dereplication using secondary metabolite gene cluster networks

Supporting Information

Theobald et al. 10.1073/pnas.XXXXXXXXXX

Supporting Information (SI)

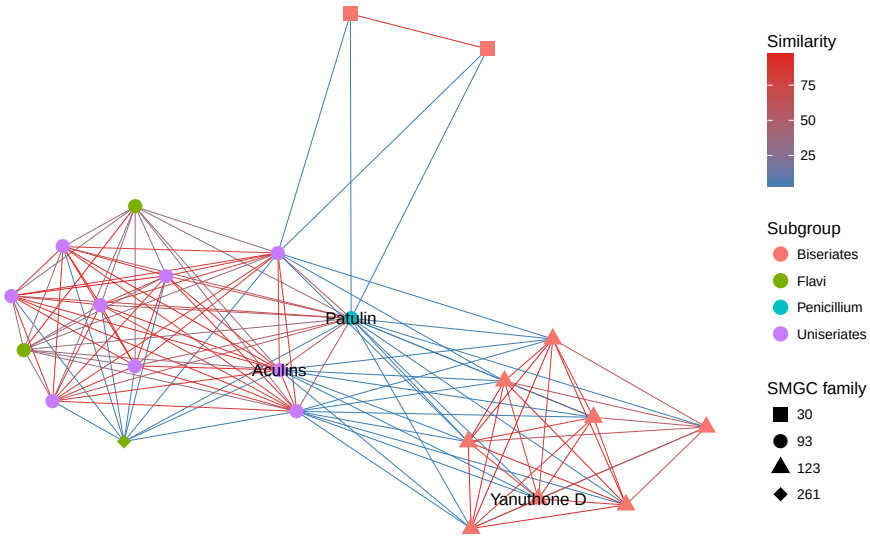


Fig. S1. Network of cluster similarities with all gene clusters connected to the patulin gene cluster. Similarity is indicated by edge color, clade by node color and annotated family by shape. The network shows four gene cluster families associated with the patulin gene cluster. Annotated gene clusters lead to the conclusion that all gene clusters in the network synthesize products based on 6-methylsalicylic acid.

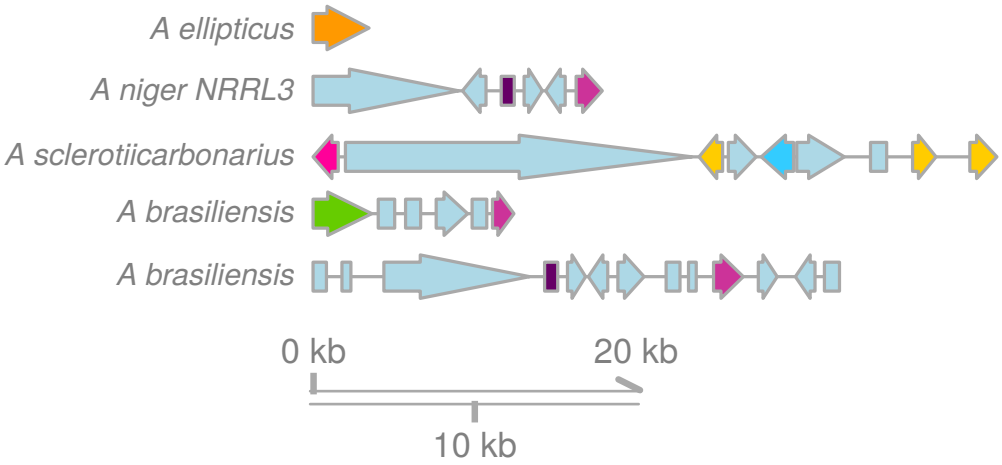


Fig. S2. Gene clusters which have been removed from predicted malformin gene cluster family. The gene clusters shown in this figure were only assigned to the family by greedy prediction or similarity to irrelevant parts of the target malformin gene cluster.

SI Figures.

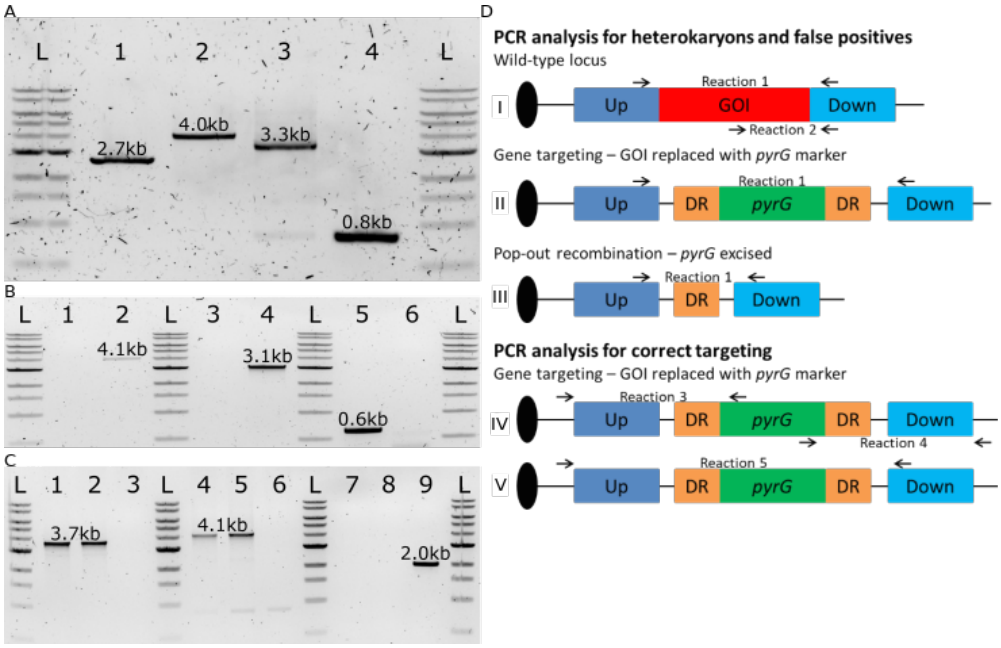


Fig. S3. A *akuA* Δ strain verification. The figure shows the PCR verification of the *akuA* (Aspbr1_0077313) deletion strain. In each lane (1-4) contains expected approximate band length, and sizes are compared to 1 kb ladder (L). Lane 1 shows reaction 1 (**D**) with primers P15+P16 (Table S2) on the wild-type strain, producing a band of 2.7 kb. After transformation of the *pyrG1* strain with *akuA* gene-targeting DNA construct, the correct integration at the *akuA* locus was tested by reaction 3 (primers P17+P18), yielding a band of 4.0 kb (lane 2), verifying the integration of the AFL*pyrG* marker into the *akuA* locus, thus, this strain is *akuA*Δ::AFL*pyrG*. Lane 3 and 4 show reaction 1 respectively on *akuA*Δ::AFL*pyrG* (3.3 kb) and the *akuA*Δ strain (0.8 kb) created after pop-out recombination of AFL*pyrG*. The marker excision band (0.8 kb) and a product of approximately 6 kb are noticeable in lane 3, which we ascribe to PCR products formed from a direct repeat annealing during PCR.
B *mlfA*Δ strain verification. The figure shows an example of the PCR verification for one of the *mlfA* (Aspbr1_34020) deletion strains (BRA30, see Table S1). The agarose gel is divided into three sections, each showing two lanes representing a specific reaction on first the wild-type strain (lanes 1, 3, 5) and secondly the *mlfA*Δ::AFL*pyrG* (lanes 2, 4, 6). The four L lanes show the 1 kb ladder. The expected approximate band length is added to lanes 2, 4 and 5, whereas in lanes 1, 3 and 6 no band is expected. The first two sections show the outcome from reaction 3 (primers P19+P24) and 4 (primers P22+P23), respectively. Lanes 1 and 3 represent PCR on the wild-type strain where AFL*pyrG* is not present in the *mlfA* locus, thus no bands are seen. Reactions on the *mlfA*Δ::AFL*pyrG* mutant (lanes 2+4) show the expected products of 4.1 kb and 3.1 kb, respectively, verifying correct targeted integration. The third and last section with lanes 5 and 6 revealed by reaction 2 (primers P20+P21) that *mlfA* was only present in wild-type strain (lane 5) but not in *mlfA*Δ::AFL*pyrG* (lane 6), where two faint unspecific smaller products was observed.
C Verification of *mlfA*-Oex strains. Lanes 1-3 and lanes 4-6 respectively show the results by agarose gel electrophoresis after PCR reaction 3 and 4 (primers P27+P28 and P29+P30), with the internal primers binding in the integrated *mlfA* instead of the *pyrG* marker. The *mlfA*-Oex strains gave the expected bands for both reaction 3 (lane 1+2; 3.7 kb) and reaction 4 (lane 4+5; 4.1 kb) confirming targeted integration of the cassette for both ends at the integration site (IS1) locus. Wild-type controls did not show any product (lane 3 and 6). Testing for presence of untransformed wild-type nuclei was accomplished by reaction 5 (primers P25+P26) in lanes 7-9, where no band is seen for the *mlfA*-Oex strains (lanes 7+8; expected band length 22kb) and a 2.0 kb band for the wild-type control (lane 9), confirming both strains to be homokaryotic mutants in the IS1 locus.
D Setup for verification of mutant strains. The PCR strategy employed ensures verification of two aspects. Firstly, the deletion cassettes with the *pyrG* marker has been correctly integrated in the targeted locus, gene of interest (GOI), and secondly, no untransformed wild-type nuclei are present in the mutant strains. All the mutant strains analyzed were compared to reference gDNA. Specifically, reaction 1 amplifies the target coding sequence from primers placed into the proximal targeting sequences (Up and Down) applied for homologous recombination. The main application of reaction 1 is to show if the strain is heterokaryon (has wild-type nuclei), **I**, while it also indicates whether the marker has replaced the target, **II**. Moreover, it is efficient to show whether the *pyrG* marker is excised after counter-selection on 5-FOA, **III**. Reaction 2 is used for very large stretches of DNA, **I**, that are difficult to amplify in reaction 1, as in the case of *mlfA*. Reactions 3 and 4 validate correct targeting by employing primers binding in unique locations outside the targeting sequences that each form a pair with primers internally in the marker gene, *IV*. Reaction 5 may likewise validate correct targeting by the one primer binding in unique location outside the targeting sequence, while also determining if the strain is heterokaryon as it provides a band even for wild-type strains since the second primer is located in the targeting sequence instead of the marker gene, **V**.

SI Tables.

Table S1. Strains used in this study

Strain ID	Genotype	Source
BRA1	Wild-type	(71)
BRA6	*pyrG1*	This study
BRA9	*pyrG1*, *akuA*Δ::AFL*pyrG*	This study
BRA10	*pyrG1*, *akuA*Δ	This study
BRA30	*pyrG1*, *akuA*Δ, *mlfA*Δ::AFL*pyrG*	This study
BRA44	*pyrG1*, *akuA*Δ, *mlfA*Δ::AFL*pyrG*	This study
BRA52	*pyrG1*, *akuA*Δ, *mlfA*Δ	This study
BRA67	*pyrG1*, *akuA*Δ, *mlfA*Δ, IS1::*PgpdA-mlfA-TtrpC*-DR-AFUM*pyrG*-DR	This study
BRA68	*pyrG1*, *akuA*Δ, *mlfA*Δ, IS1::*PgpdA-mlfA-TtrpC*-DR-AFUM*pyrG*-DR	This study

Nomenclature for the TSs refers to the numbered species in the table. AFUM: *Aspergillus fumigatus*, AFL: *A. flavus*

Table S2. List of primers used in this study

Code	Name	Sequence (5'-3')
P1	gRNA-ABRApyrG-rv	AGCTTACUCGTTTCGTCCTCACGGACTCATCA-GGCCTCTCGGTGATGTCTGCTCAAGCG
P2	gRNA-ABRApyrG-fwd	AGTAAGCUCGTCGCCTCTTGGCAAGAGCATTGG-TTTTAGAGCTAGAAATAGCAAGTTAAA
P3	gRNA-ABRAakuA-rv	AGCTTACUCGTTTCGTCCTCACGGACTCATCAG-GATGAACGGTGATGTCTGCTCAAGCG
P4	gRNA-ABRAakuA-fwd	AGTAAGCUCGTCGATGAAGATGAAGACAGTCG-GTTTTAGAGCTAGAAATAGCAAGTTAAA
P5	PgpdA-pac-up-fwd	GGGTTTAAUGCGTAAGCTCCCTAATTGGC
P6	TtrpC-short-pac-dw-rv	GGTCTTAAUGAGCCAAGAGCGGATTCCTC
P7	ABRAakuA-Dl-Up-FU	GGGTTTAAUGACTGGTGAGTGATTGTGGGAG
P8	ABRAakuA-Dl-Up-RU	GGACTTAAUGTGGGGCTGCTGTGTCTG
P9	ABRAakuA-Dl-Dw-FU	GGCATTAAUGCATAATAGGATGGTCGGGTTCG
P10	ABRAakuA-Dl-Dw-RU	GGTCTTAAUGCCTTGGTAGGCAGACGAATAG
P11	ABRA34020-Dl-Up-FU	GGGTTTAAUGAGGGAGTGAAAGTCGTCG
P12	ABRA34020-Dl-Up-RU	GGACTTAAUGAAGGAGGTGAAGTTACTAGGAC
P13	ABRA34020-Dl-Dw-FU	GGCATTAAUGTGGTTGCCATGAGTGAAAGTG
P14	ABRA34020-Dl-Dw-RU	GGTCTTAAUGATGCAGTGCAGGTTGGAG
P15	ABRAakuA-ChkGap-F	CCAGCCAGCGTCATCAATTAC
P16	ABRAakuA-ChkGap-R	CCTACCCCGACATCCAACC
P17	ABRAakuA-ChkUP-F	CCTTCCCAGCTCTCAAGTCC
P18	AFLpyrG-Chk-int-R2	CTAGATCACATGTAAGTGGCATCCC
P19	ABRA34020-Chk-Up-F	CCGTCCGAAGATCAATCCGAC
P20	ABRA34020-Chk-Gap-F	GCACGTCGGCTGTGATGG
P21	ABRA34020-Chk-Int5'-R	GAGTGTGATCTGGATTCGGG
P22	ABRA-34020-ChkDw-R	CCCACCATTTGTAACGCACTG
P23	AFLpyrG-Chk-Dw-F	CCCACCACCCCCTACTCTAACAC
P24	AFLpyrG-Int-R-BP	CCCATCACAAACTTCTTATACTTCCGA
P25	ABIS3-ChkDw-R	CACCACAGCCATCAAATCC
P26	ABIS3-ChkGap-F	CGACAACCCCAACAACG
P27	ABIS3-Chk-Up-F2	GGTGTCTGTGTGTGACAGACTAG
P28	ABRA34020-Chk-Int5'-R	GAGTGTGATCTGGATTCGGG
P29	ABIS3-Chk-Dw-R2	GTGTCTTTGGTACCATGGGAGAG
P30	BRAmalA-GapChk-F	GAAGTGTGGGGTTGATGTG
P31	ABRA-IS-Up-FU2	GGGTTTAAUGGCTTCTCCATCCTTCTACATG
P32	ABRA-IS-Up-RU2	AGGGAAUTGTAGGTAAACCAGCACGCTC
P33	ABRA-IS-Dw-FU	GGCATTAAUGATGAGCATATTGTGTGAGG
P34	ABRA-IS-Dw-RU	GGTCTTAAUGGGATGGGTATGTTGGA
P35	PgpdA(2.3kb)-FU-ANIS5-TSI	ATTCCCUTGTATCTCTACACACAGGC
P36	PgpdA-Rev-RU PacI.Reg	AAACCCUCAGCGCGGTAGTGATGTCTGCTCAAG
P37	TtrpC-PacI.Reg-FU	AGGGTTUAATTAAGACCTCAGCGGATCCACTTAACGTTAC
P38	TtrpC-RU-PacI	GGACTTAAUCGCTTACACAGTACACGAG
P39	ABRA34020-F1-FU-PacI 2	GGGTTTAAUATGAGTCGCTTTTCCTGC
P40	ABRA34020-F1-RU 2	AGCCCCTGUCCGTGTGTT
P41	ABRA34020-F2-FU	ACAGGGGCUGTTCAACACG
P42	ABRA34020-F2-RU	AGCGAGAUGCGGACAGG
P43	ABRA34020-F3-FU 2	ATCTCGCUCGGTGACCAAG
P44	ABRA34020-F3-RU 2	ACTCCGUCAACTGGAACATCTC
P45	ABRA34020-F4-FU 2	ACGGAGUGGAGATGATGGTC
P46	ABRA34020-F4-RU-PacI	GGTCTTAAUTCAAACACAGACACCCCGAG

P1-6 CRISPR/*cas9* construction, P7-14 deletion plasmid construction, P15-30 strain verification. P31-46 Integration site (IS) and *mlfA* complementation.

Table S3. List of plasmids used in this study

Plasmid ID	Application	Content
pFC334 (44)	CRISPR sgRNA template	*argB*, *Ptef1-cas9-Ttef1*, *PgpdA*-sgRNA-*TtrpC*
pFC332 (44)	CRISPR/*cas9* vector	*hygR*, *Ptef1-cas9-Ttef1*
pFC837	CRISPR/*cas9 pyrG*[b]	*hygR*, *Ptef1-cas9-Ttef1*, *PgpdA*-sgRNA(*pyrG*)-*TtrpC*
pFC715	CRISPR/*cas9 akuA*[c]	*hygR*, *Ptef1-cas9-Ttef1*, *PgpdA*-sgRNA(*akuA*)-*TtrpC*
pFC478 (44)	Gene deletion vector	DR-AFL*pyrG*-DR
pFC645	Deletion plasmid *akuA*	Up_{akuA}-DR-AFL*pyrG*-DR-$Down_{akuA}$
pFC809	Deletion plasmid *mlfA*	Up_{mlfA}-DR-AFL*pyrG*-DR-$Down_{mlfA}$
pFC3 (43)	Base overexpression vector	DR-AFUM*pyrG*-DR
pFC1116	*mlfA* overexpression	Up_{IS1}-*PgpdA-mlfA-TtrpC*-DR-AFUM*pyrG*-DR-$Down_{IS1}$

b: Protospacer for targeting *A. brasiliensis pyrG* is GCCTCTTGGCAAGAGCATTG, c: Protospacer for targeting *A. brasiliensis akuA* is GATGAAGATGAAGACAGTCG. All plasmids contain the *ampR* and *AMA1* sequence for propagation in *E. coli* and Aspergilli, respectively.

A.3 Supplementary information: Genus level analysis of PKS-NRPS and NRPS-PKS hybrids reveals their origin in Aspergilli

Supplementary information for: Genus level analysis of PKS-NRPS and NRPS-PKS hybrids reveals their origin in Aspergilli

January 31, 2018

152

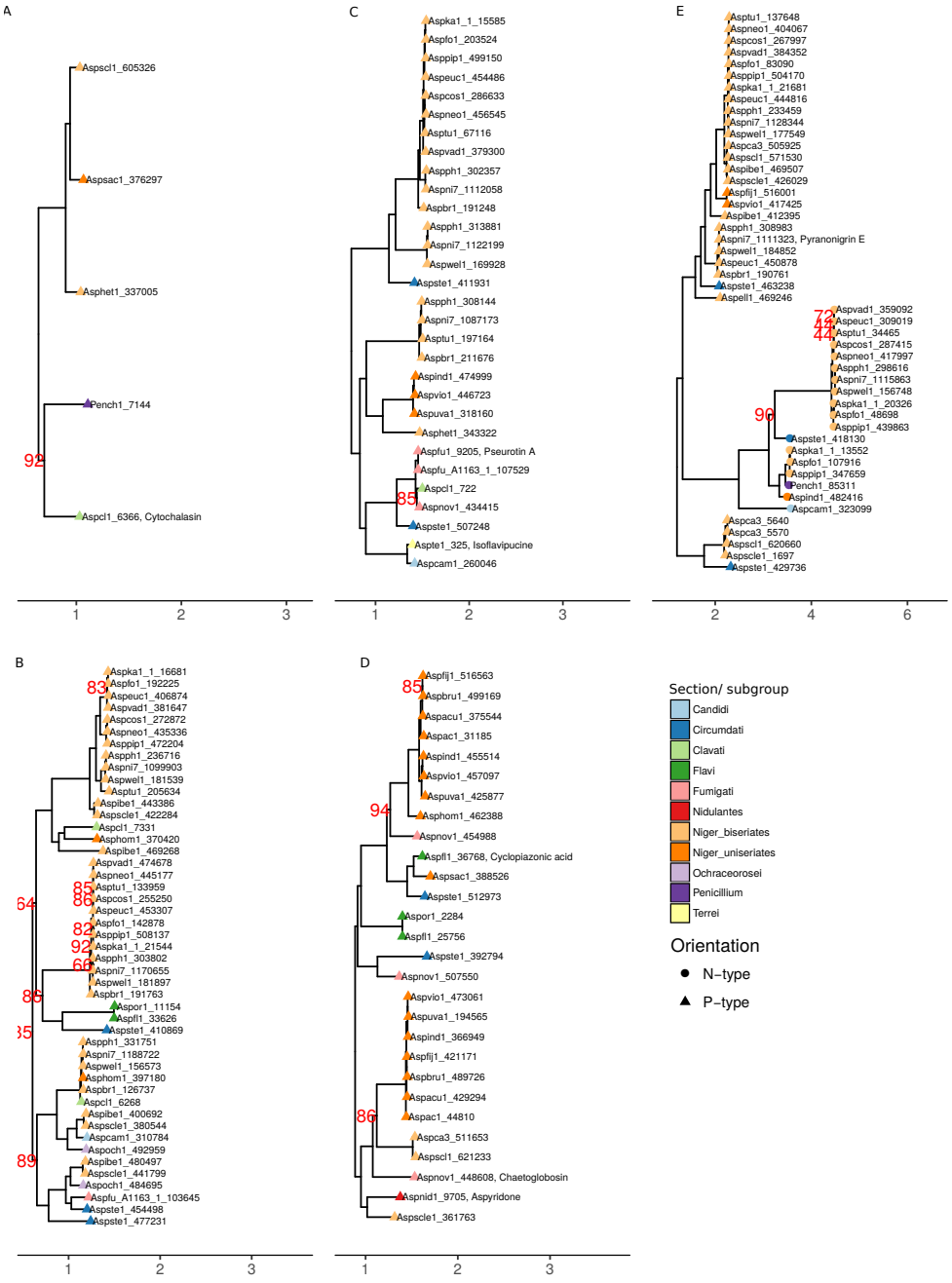


Figure 1: Selected nodes from hybrid maximum liklihood phylogeny. Subtrees were extracted from hybrid tree figure 1. Sections and species groups indicated by tip color; Orientation of hybrids N-type (NRPS-PKS) and P-type (PKS-NRPS) indicated by tip shape. Tip labels constist of jgi organism identifier, jgi protein id and associated compound (if applicable).

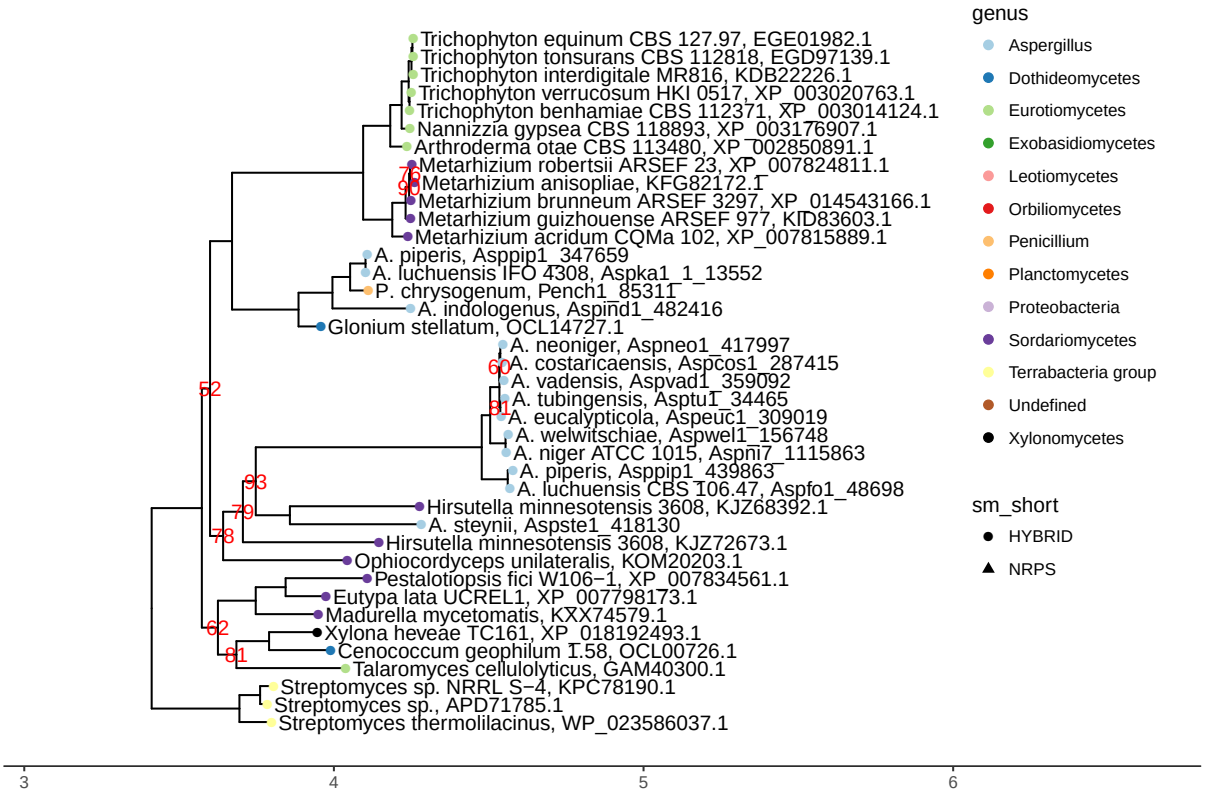


Figure 2: ML phylogeny of fungal and bacterial hybrids. Subtree extracted from figure 6. Tip labels show species name and NCBI identifier/ JGI organism and protein identifier. Tip color indicates genus or class; Tip shape indicates SM protein type. Bootstrap values under 95 are shown in red. Hybrids from *Streptomyces* form a sister clade to fungal hybrids NRPS-PKS hybrids, indicating this class from bacterial origin.

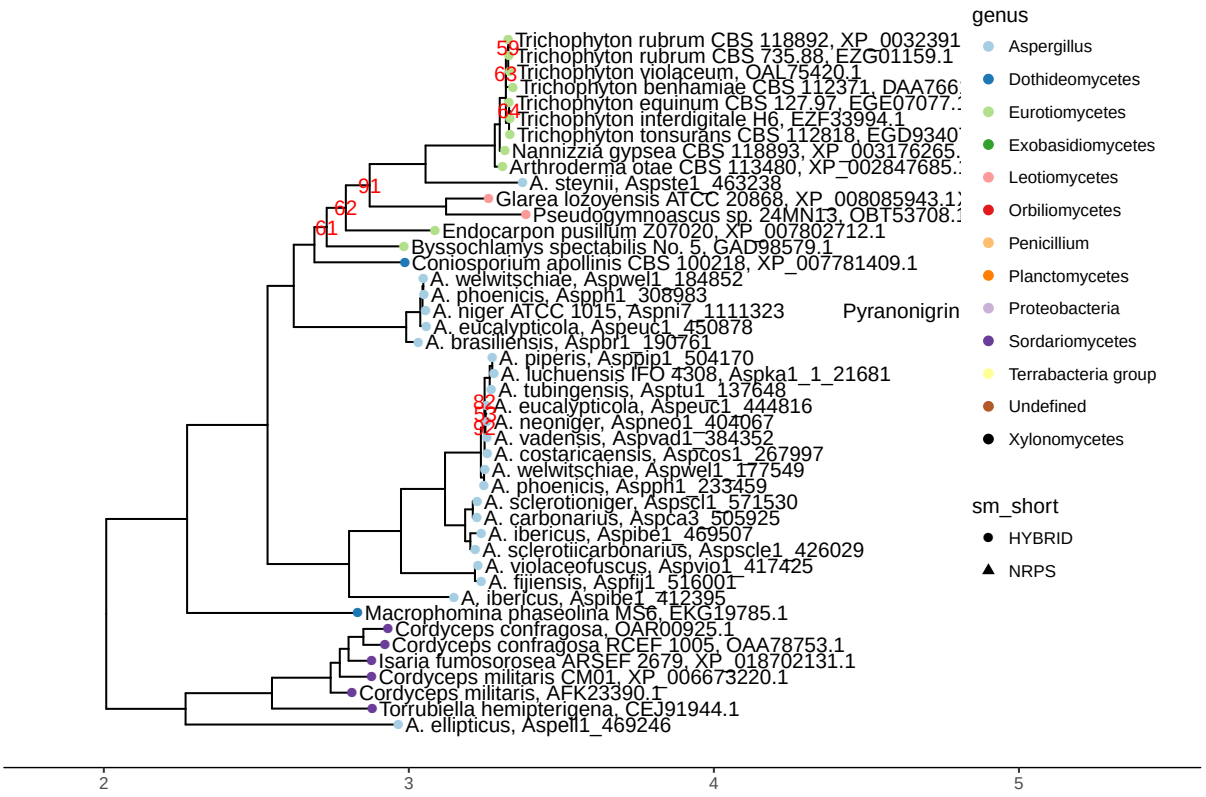


Figure 3: ML phylogeny of pyranonigrin associated hybrids. Subtree extracted from figure 6. Tip labels show species name and NCBI identifier/ jgi organism and protein identifier. Tip color indicates genus or class; Tip shape indicates SM protein type. Bootstrap values under 95 are shown in red. Additional tip label shows associated compound (if applicable).

A.4 Supplementary information: Comparative genomics of *A. nidulans* and section *Nidulantes*

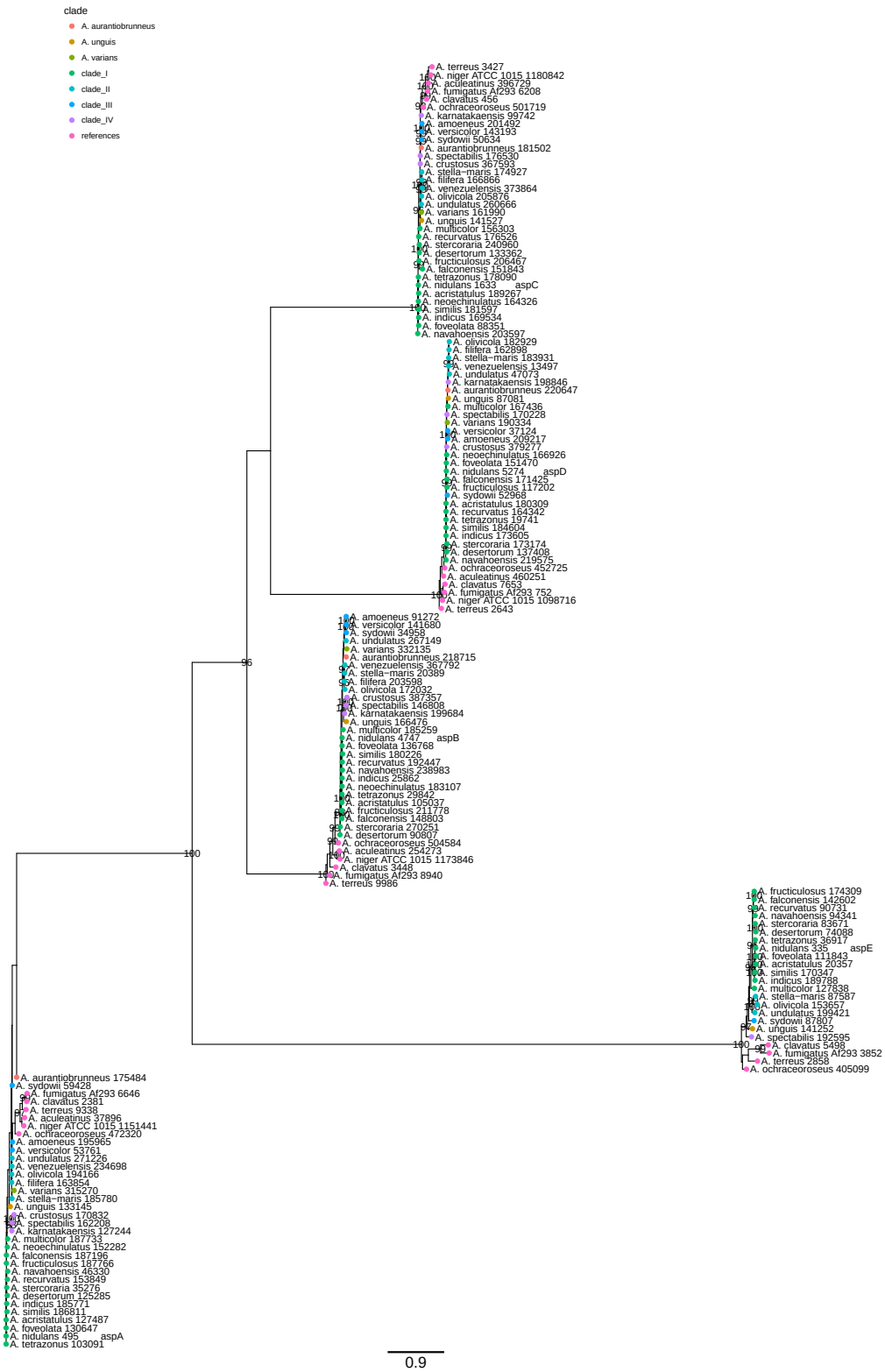


Fig. S1. Maximum likelihood phylogeny of septin homologues

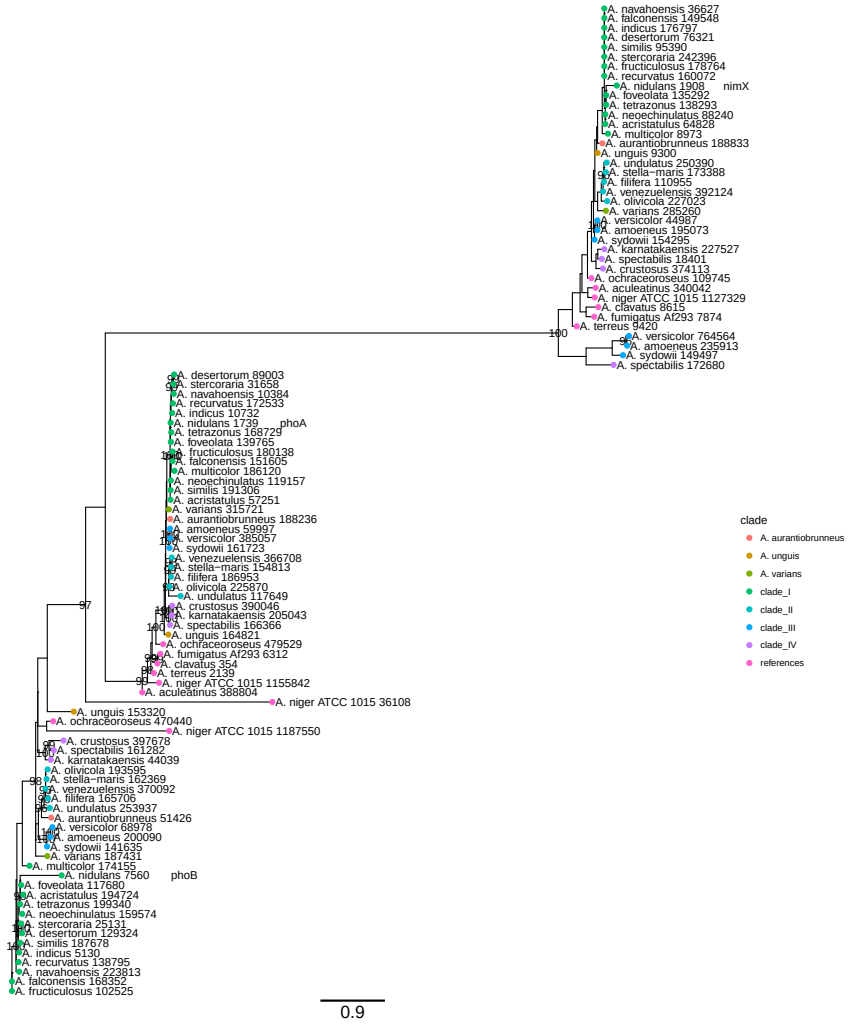
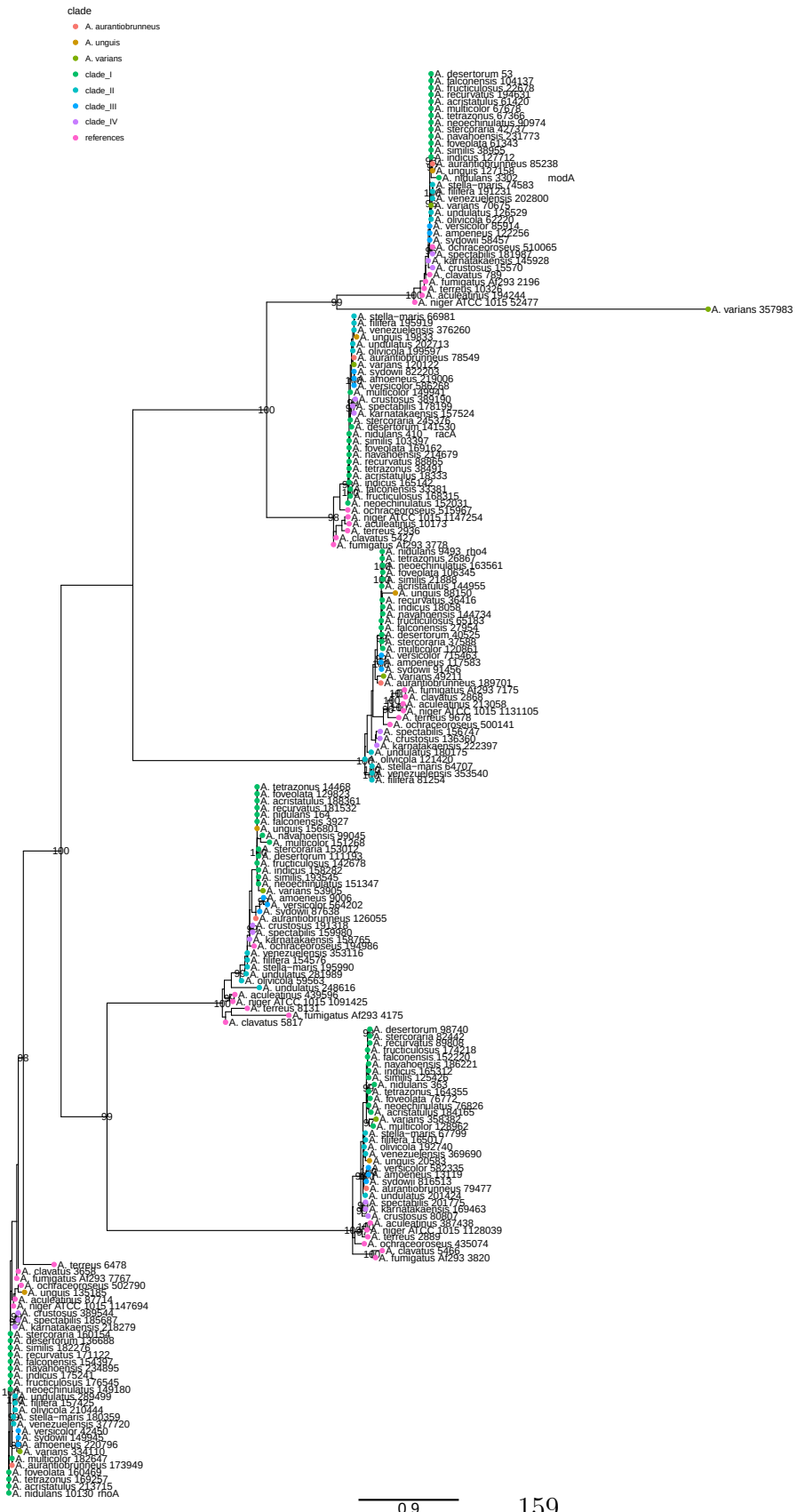


Fig. S2. Maximum likelihood phylogeny of nimX, phoA, phoB homologues


clade
A. aurantiobrunneus
A. unguis
A. varians
clade_I
clade_II
clade_III
clade_IV
references
A. varians 357983
modA
racA
rho4
0.9

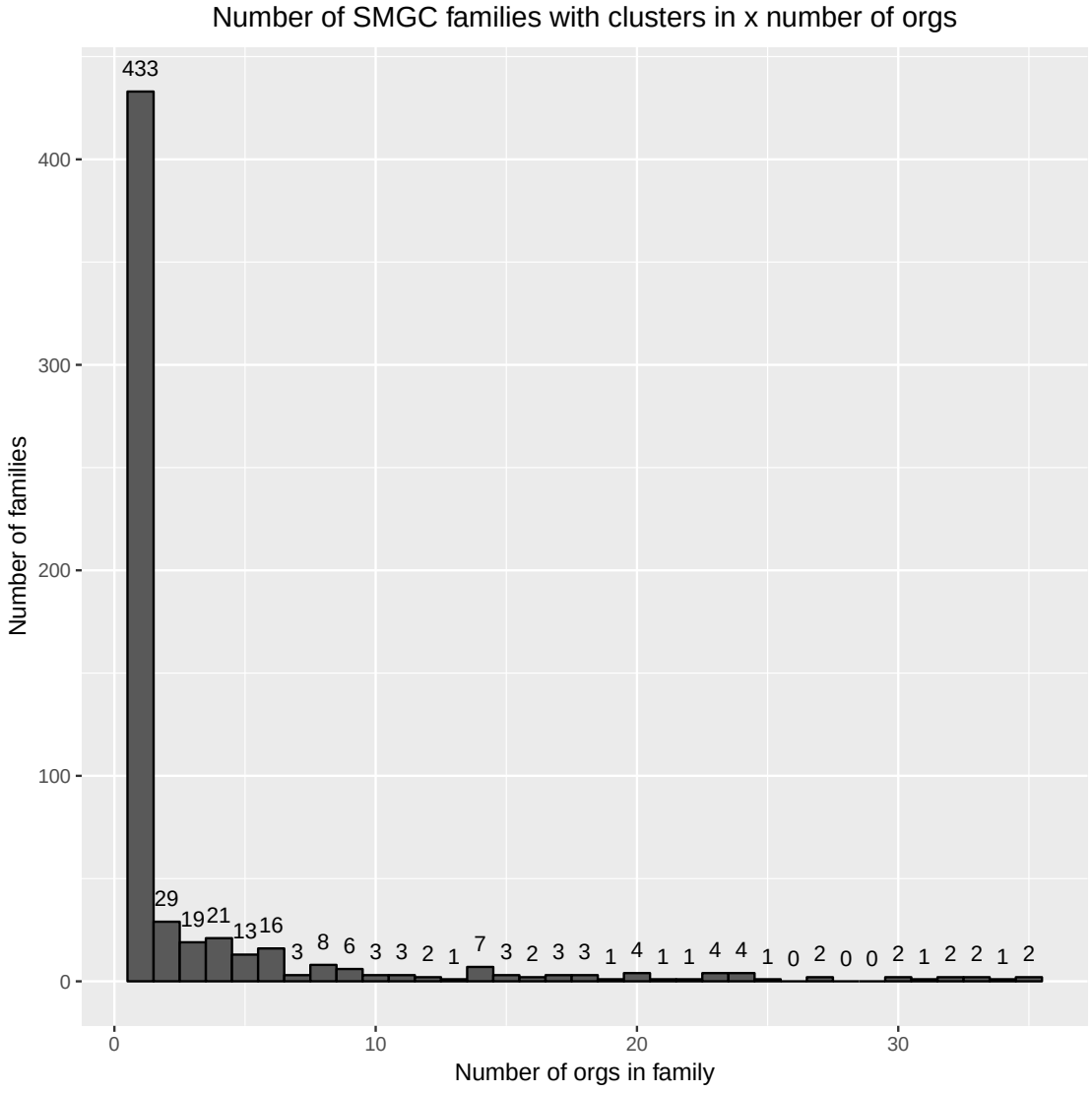


Fig. S4. SMGC family counts by size. Barplot showing counts of families with certain number of organisms.

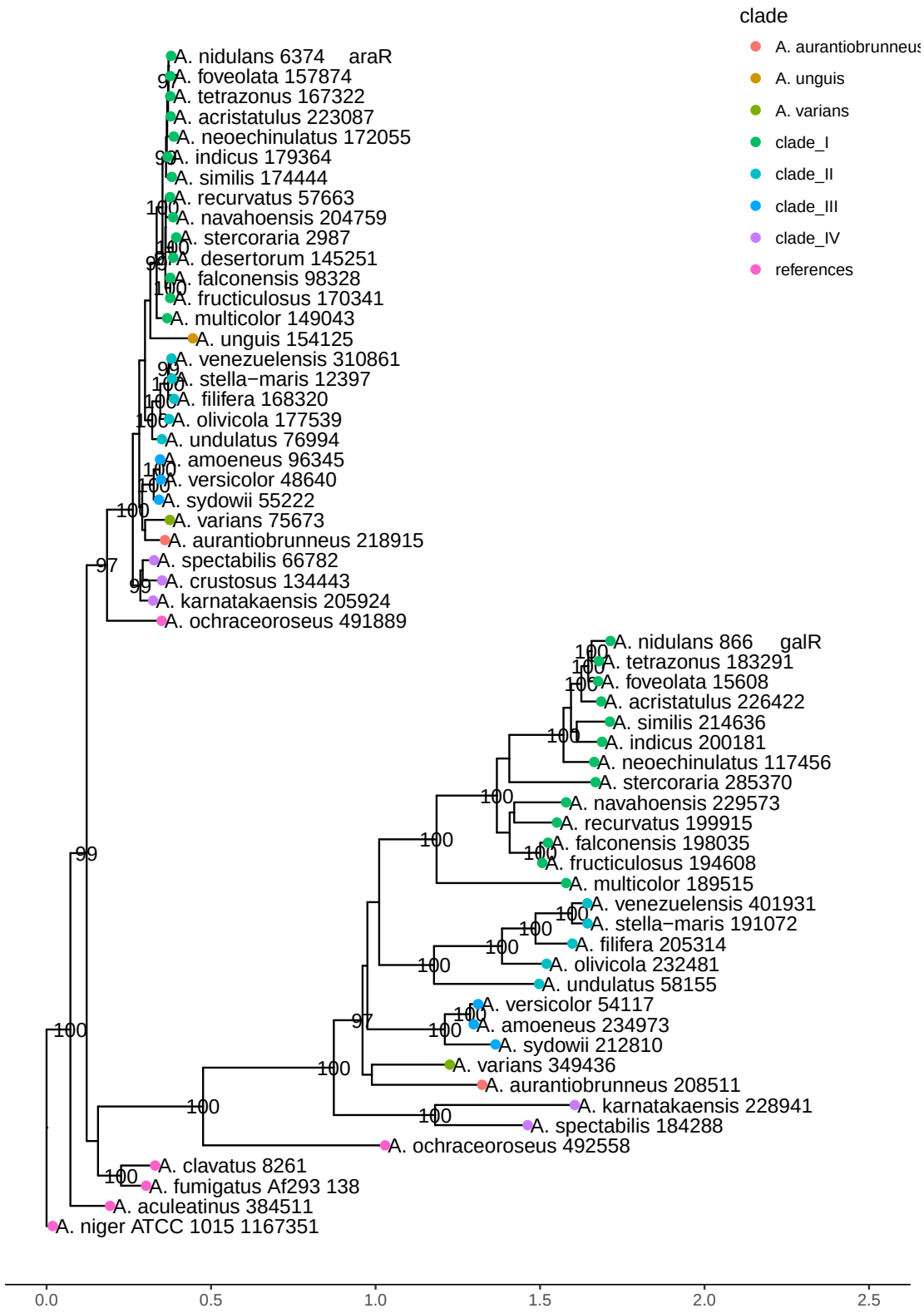


Fig. S5. Maximum likelihood phylogeny of protein family containing *A. nidulans araR* and *galR*. Tip labels are showing species name, jgi protein id and gene name (if applicable). Proteins were aligned using clustalo (74), trimmed using trimal (75) and a ML phylogeny created using iqtree (76) using a WAG+F+I+G4 substitution model. The ML phylogeny suggests a galR homolog in all species of section *Nidulantes* and potentially *A. ochraceroseus*.

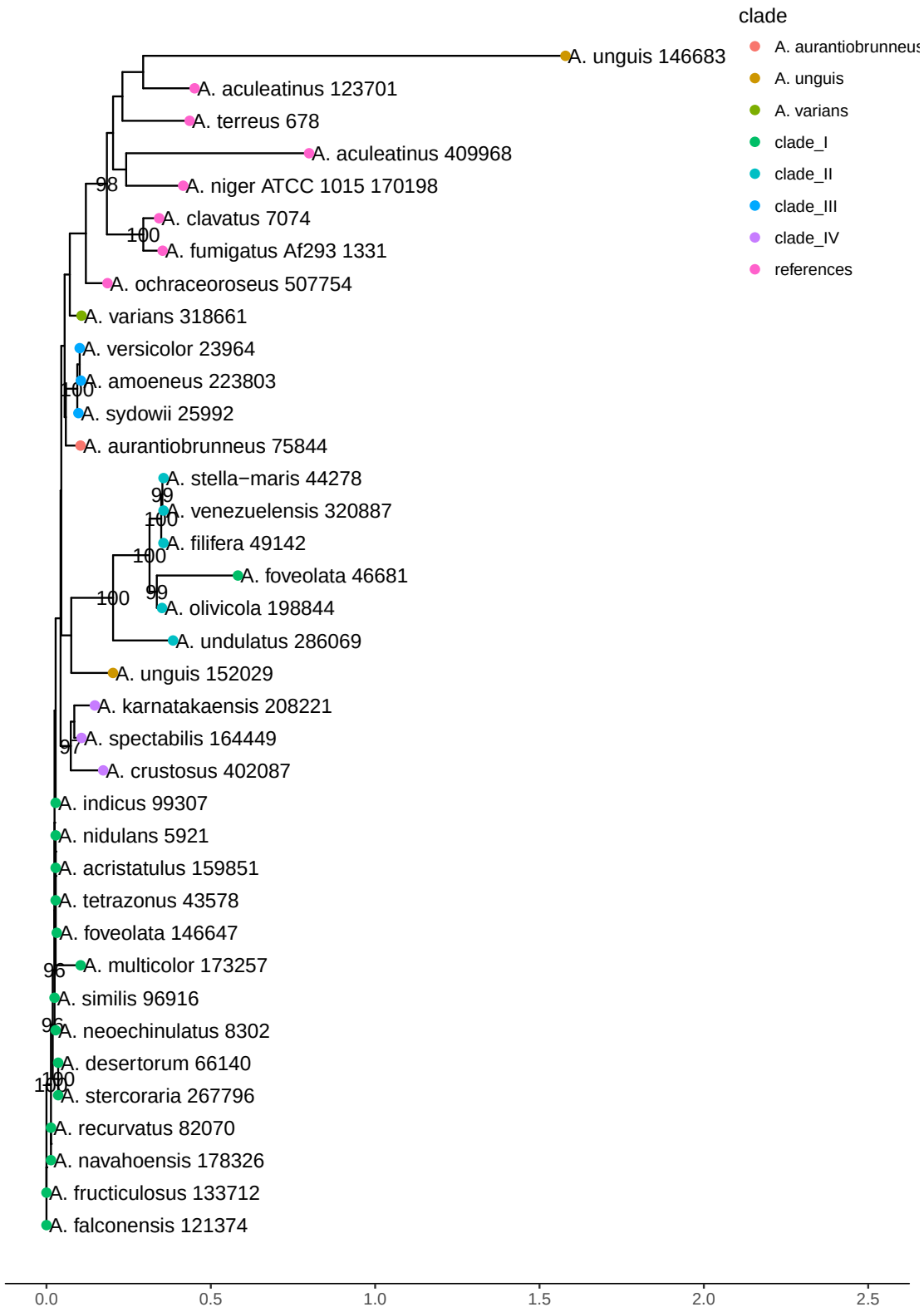


Fig. S6. Maximum likelihood phylogeny of laeA protein family. Tip labels are showing species name and jgi protein id. Proteins were aligned using clustalo (74), trimmed using trimal (75) and a ML phylogeny created using iqtree (76) using a LG+G4 substitution model (ultra fast bootstrapping (77), model finder plus (78)). Species show one copy except *A. foveolata*, *A. aculeatinus* and *A. unguis*, which contain another copy.

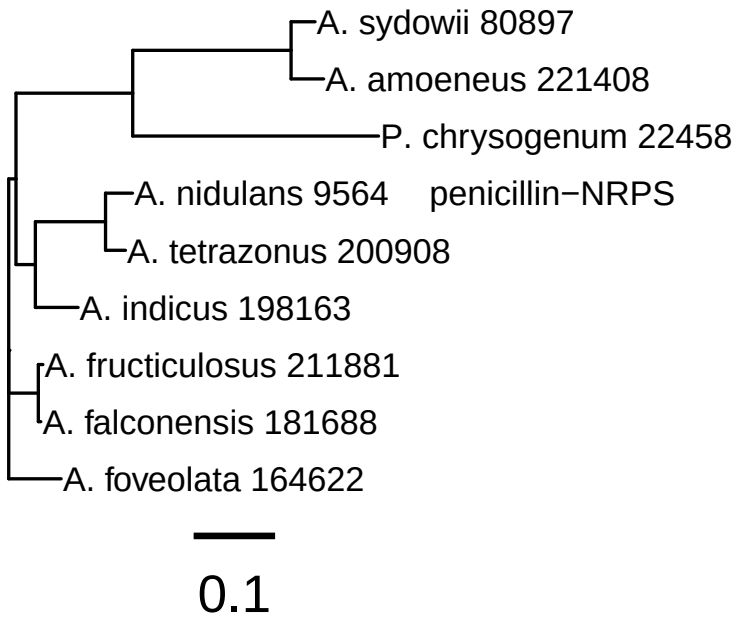


Fig. S7. ML phylogeny of penicillin NRPS homologs. Proteins were aligned using clustalo (74) and ML phylogeny created using iqtree (76) with a WAG+F+G4 substitution model (ultra fast bootstrapping (77), model finder plus (78))

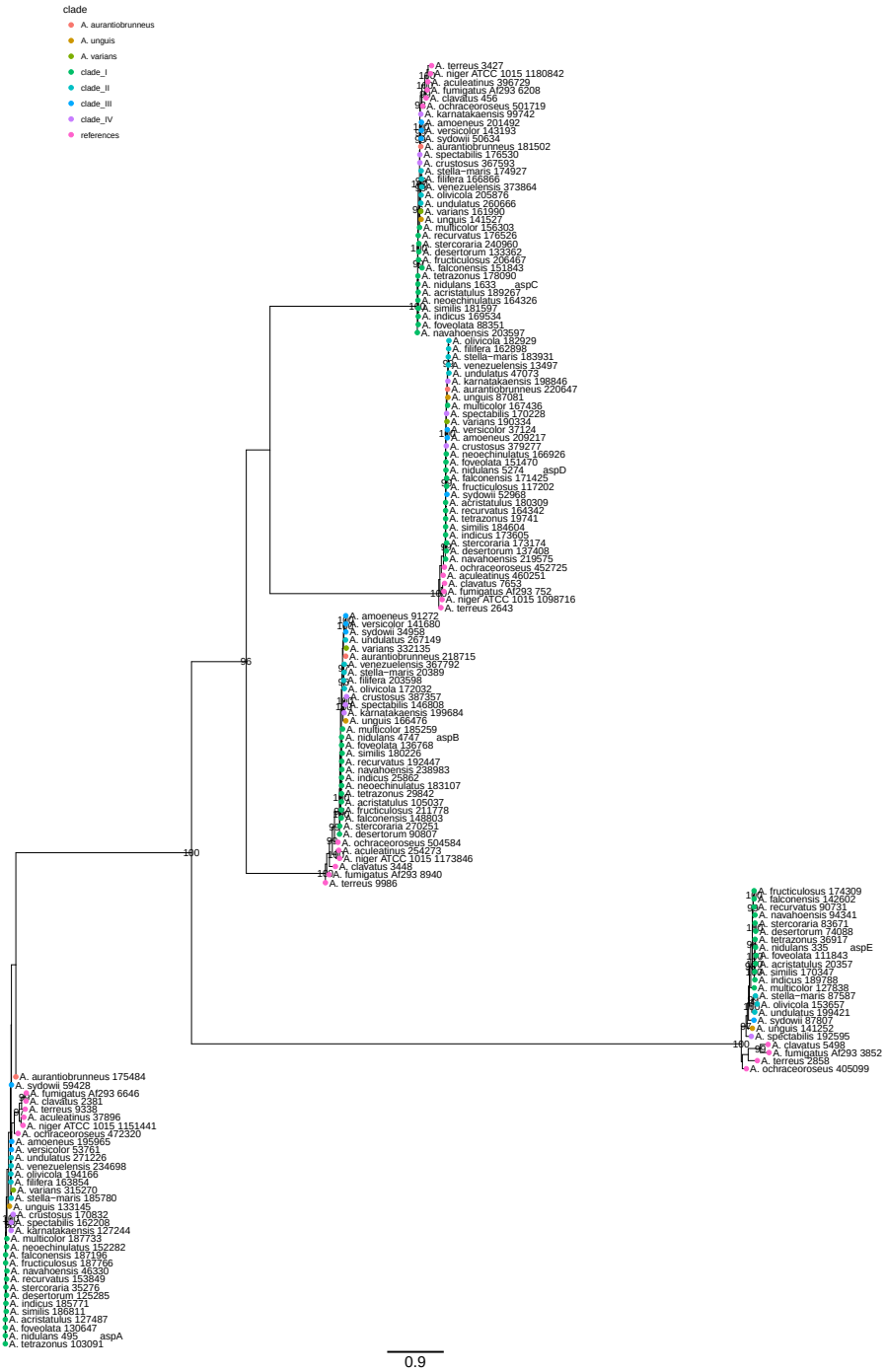


Fig. S8. Maximum likelihood phylogeny of septin protein family.

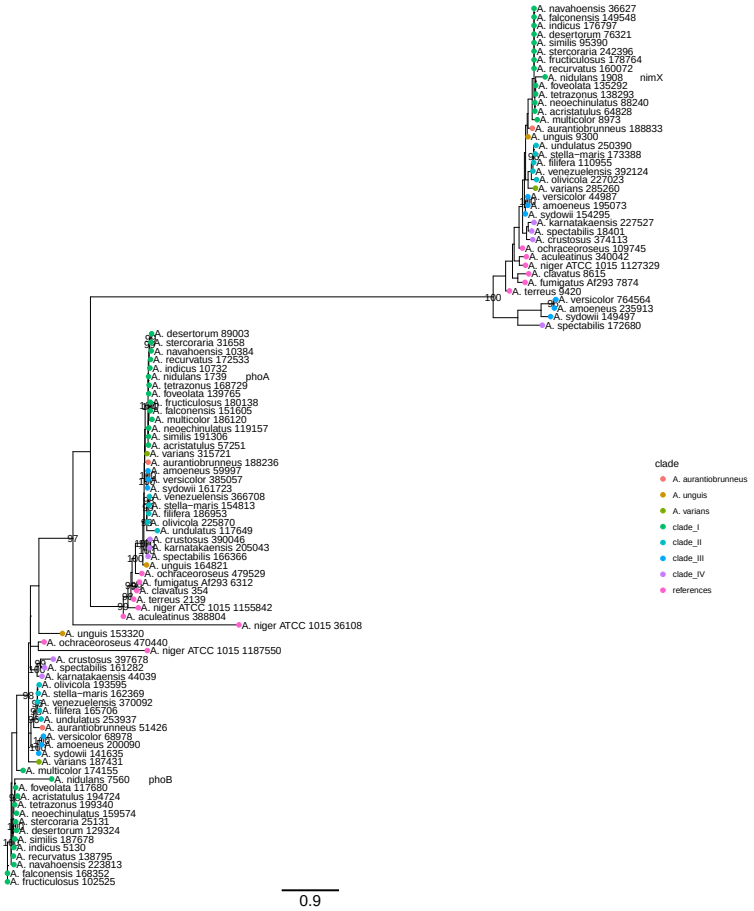


Fig. S9. Maximum likelihood phylogeny of nimX, phoA, phoB protein family.

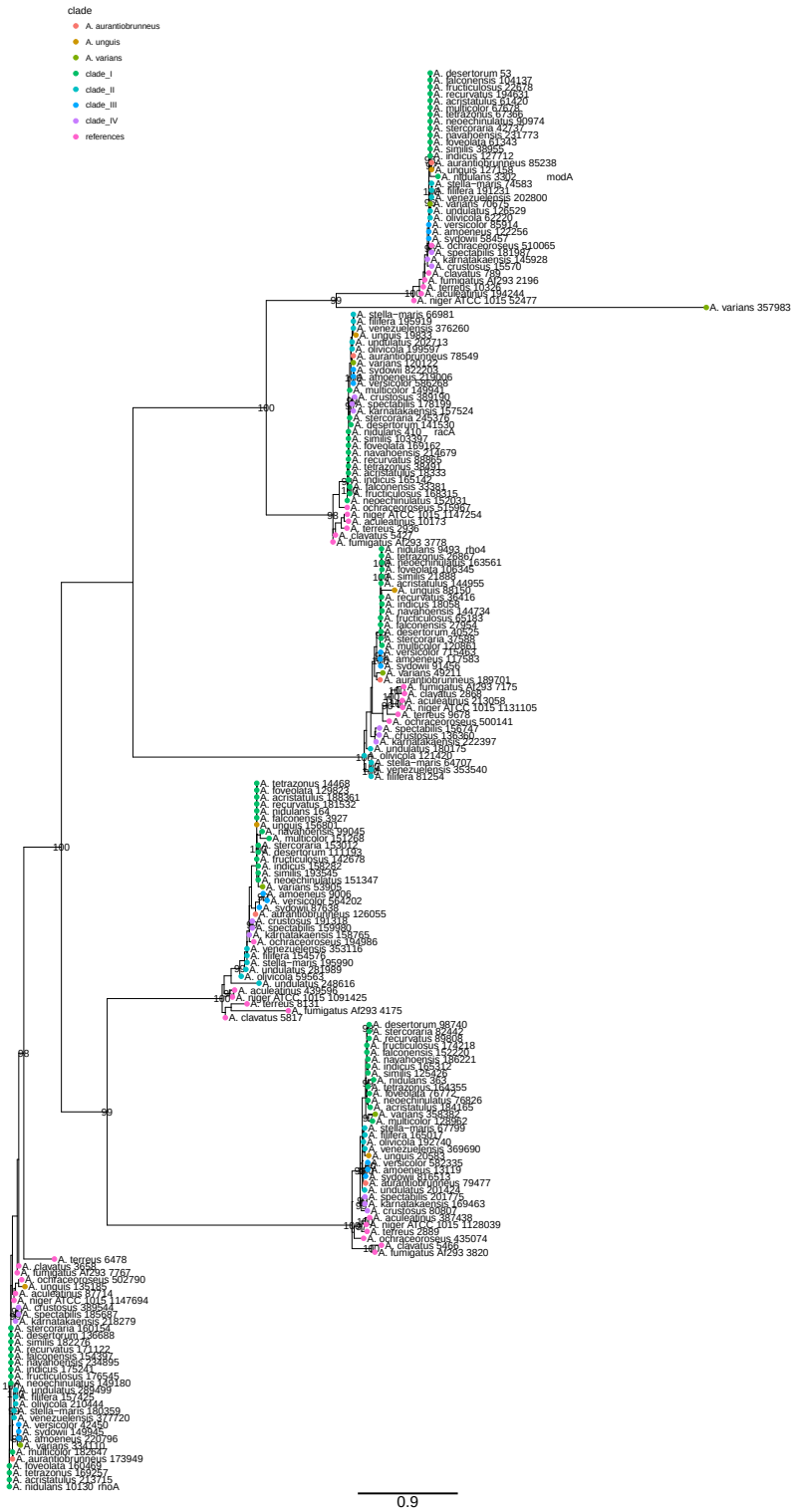


Fig. S10. Maximum likelihood phylogeny of rho4, racA, modA protein family.